University of Alberta

Development of Mass Spectrometric Methods for Rapid and Sensitive Bacterial Identification

by

Zhengping Wang ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment
of the requirements for the degree of Doctor of Philosophy

Department of Chemistry

Edmonton, Alberta

Fall 2001

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-69016-4

Canada

University of Alberta

Library Release Form

Name of Author: Zhengping Wang

Title of Thesis: Development of Mass Spectrometric Methods for Rapid and Sensitive Bacterial Identification

Degree: Doctor of Philosophy

Year this Degree Granted: 2001

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Zhengping Wang

602B, Michener Park
Edmonton, AB, Canada, T6H 5A1

September 26, 2001

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled Development of Mass Spectrometric Methods for Rapid and Sensitive Bacteria Identification submitted by Zhengping Wang in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Liang Li
Dr. L. Li, Professor of Chemistry

Dr. C. A. Lucy, Professor of Chemistry

Dr. J. B. D. Green, Assistant Professor of Chemistry

Monica Palcic
Dr. M. M. Palcic, Professor of Chemistry

P Fedorak
Dr. P. M. Fedorak, Professor of Biological Sciences

CAL
Dr. D. M. Lubman, Professor of Chemistry, University of Michigan

Sept. 26, 2001

To my parents, my husband and my lovely daughter

Abstract

Mass spectrometry analysis of proteins or a subset of a proteome (i.e., proteins expressed by the genome of a cell) is a potentially very sensitive and specific approach for unambiguous bacterial identification. This work focuses on developing methods to improve the matrix assisted laser desorption/ionization (MALDI) technique for bacterial protein analysis, and understanding issues related to the creation of protein mass database tailored for bacterial identification.

Bacterial protein mass spectral pattern obtained by direct MALDI analysis of crude cell extracts is very sensitive to minor experimental condition changes. Different subsets of bacterial proteome may be detected under different sample extraction/preparation conditions. Therefore, it is not reliable to differentiate bacteria on the basis of their different mass spectral patterns. Alternatively, bacterial identification can be achieved by searching a set of protein masses against a bacterial protein mass database. This approach will not be affected by the variations in the set of protein masses observed under different conditions, since the sets of protein masses should always reflect the bacterial genome.

A confident bacterial identification greatly relies on the quality of the set of protein masses obtained by direct MALDI analysis of the crude cell extracts. Optimization of protein extraction and MALDI sample preparation to detect a large number of proteins in a broader mass window is demonstrated in this work.

The availability of a comprehensive and reliable bacterial protein mass database is also very critical for an unambiguous bacterial identification. The protein masses in the public proteome database are mostly derived from their genome translated protein

sequences. In reality, most proteins have involved in some kinds of processing after translation. Proteins can also be modified or processed *in vitro* during sample preparation. As such, proteome database cannot be directly used for the purpose of bacterial identification. A bacterial protein mass database specifically tailored for bacterial identification can be easily created by MS methods. Preliminary results have demonstrated that such a database is potentially very useful for bacterial identification.

Acknowledgements

First of all, I would like to thank my supervisor Dr. Liang Li for his support, encouragement and his excellent guidance throughout my study and research. Besides his expertise in mass spectrometry and related areas, his dedication and enthusiasm towards our research goal have greatly inspired me. I appreciate the chance working in his research group and I am sure what I have learned here will be invaluable for my future career.

I also thank all my committee members, Dr. Charles A. Lucy, Dr. Monica M. Palcic, Dr. John-Bruce. D. Green, Dr. Phillip M. Fedorak and Dr. David M. Lubman for their patience and very helpful suggestions to my thesis.

I would like to give my special thanks to Dr. Norman J. Dovichi, who used to be in my committee and is now in University of Washington, I am very grateful to him for his encouragement during the early stage of my Ph. D. study.

My deep gratitude also goes to all the members in Dr. Liang Li's research group. I would like to thank Dr. Bernd O. Keller, Mr. Alan Doucette, Mrs. Jing Zheng and Mr. Kevin Dunlop for the co-operation in different projects. I would like to thank Mrs. Lidan Tao for providing bacterial samples. I thank Mr. Alan Doucette for spending his precious time proofread my thesis. I thank Dr. Truong Ta, Mr. Rui Chen, Mrs. Nan Zhang and Mr. Chris McDonald for their kindness help and very useful discussion when I was writing my thesis and preparing my defense. I also would like to thank our former group members, Dr. Wojciech Gabryelski, Dr. Randy Whittal, Dr. Yuqin Dai and Dr. Ken K.-C. Yeung, for their advice and very helpful discussion.

I thank all the support from Department of Chemistry, University of Alberta. Especially the personnel in general, purchasing and post offices, the machine and electronic shop and the store room, I will never forget their kindness help.

I would like to thank all the friends who have given me so much support and encouragement. I specially thank Rong Jiang and Yong Gao for their friendship, advice and great help during the first year of my Ph. D study.

Finally, I would like to thank my parents, brothers and sister for always being so supportive and understanding. I would also like to thank my parents-in law for all the help they gave to me, without whom I would not be able to concentrate on my study and research in the past year. I thank my husband, Linfeng Fan and my daughter, Yujing Fan for their love, support, most importantly for filling my life with great joy.

Table of Contents

Chapter 1 1

Introduction to Bacterial identification by Mass Spectrometric Techniques..........1

1.1 Bacterial identification by Mass Spectrometry Using Proteins as Biomarkers 4

1.2 Mass Spectrometry 7

1.2.1 MALDI Time-of-Flight (TOF) MS 7

1.2.2 ESI-Quadrupole and ESI Ion Trap 14

1.3 Protein Identification by Mass Spectrometry..........17

1.3.1 Peptide Mass Mapping 18

1.3.2 Short Protein/Peptide Sequence Tag 19

1.3.3 Peptide Fragment Ion Fingerprinting by MS/MS 20

1.4 Brief Summary of the Thesis 22

1.5 Literatures Cited 23

Chapter 2 29

Investigation of Spectral Reproducibility in Direct Analysis of Bacteria Proteins by MALDI-TOFMS 29

2.1 Introduction 29

2.2 Experimental 30

2.2.1 Chemicals and Materials 30

2.2.2 Extraction of Bacterial Proteins 30

2.2.3 MALDI analysis 31

2.3 Results and Discussion 32

2.3.1 Effect of the Sample Solvent 32

2.3.2 Effect of the Salt Content 35

2.3.3 Effect of the Extraction Solvent 38

2.3.4 Effect of the Extraction Method 44

2.3.5 Inter-laboratory Comparison 47

2.3.6 Comparison with Related Efforts 50

2.4 Conclusions 52
2.5 Literatures Cited 55

Chapter 3 57
Mass Spectrometric Methods for Generation of Low-Mass Proteome Database to be Used for Bacterial identification 57
3.1 Introduction 57
3.2 Experimental 58
3.2.1 Bacterial Protein Extraction 58
3.2.2 HPLC Fractionation 58
3.2.3 Mass Spectrometry 59
3.2.3.1 ESI..59
3.2.3.2 MALDI Analysis 60
3.3 Results 61
3.3.1 LC/ESI-MS 61
3.3.2 Direct MALDI 63
3.3.3 LC/Off-line MALDI 63
3.4 Discussion 68
3.4.1 Public Proteome Database 68
3.4.2 Comparison of Three Data Sets 71
3.4.2 Comparison of MS Data with Published Proteome Database......73
3.4.3 Evaluation of the Applicability of Protein Mass Tables for Bacterial identification 74
3.4.4 Mass Database Creation 77
3.5 Conclusions..78
3.6 Literatures Cited.. 79

Chapter 4 81
Matrix-Assisted Laser Desorption Ionization Mass Spectrometry and Gel Electrophoresis Analysis of Bacterial Proteome from Rapid Solvent Extraction of Bacterial Cells 81
4.1 Introduction 81

4.2 Experimental 83
4.2.1 Materials 83
4.2.2 Bacterial Protein Extraction 83
4.2.3 Total Protein Determination 83
4.2.4 SDS-PAGE 84
4.2.5 Extraction of Intact Proteins from the Gel 84
4.2.6 In-Gel Digestion 84
4.2.7 Mass Spectrometry 85
4.3 Results and Discussion 87
4.4 Conclusions 102
4.5 Literatures Cited 104

Chapter 5..........107
Identification of Low Mass Bacterial Proteins and Their Post-translational Modifications by HPLC Separation, Enzymatic Digestion and MALDI 107
5.1 Introduction 107
5.2 Experimental 109
5.3 Results and Discussion 110
5.3.1 Identification of *E. coli* Proteins and Their Post-translational Modifications by Tryptic and Chymotryptic Peptide Mass Mapping 110
5.3.2 N-terminal Digestion Using Leucine Aminopeptidase (LAP)....118
5.4 Conclusions 122
5.5 Literatures Cited 123

Chapter 6 126
Nanoliter Protein Concentration and Digestion Combined with Microspot MALDI MS for Identification of Proteins Fractionated by Conventional HPLC 126
6.1 Introduction 126
6.2 Experimental 127
6.2.1 Chemicals and Materials 127

6.2.2 Extraction of Bacterial Proteins 128
6.2.3 HPLC Fractionation 128
6.2.4 In-capillary Sample Concentration, Reaction and Microspot MALDI Sample Preparation 128
6.2.5 MALDI Analysis 132
6.3 Results and Discussion 132
6.3.1 Effects of In-capillary Concentration and Cleaning Steps......132
6.3.2 Bacterial Protein Identification by In–capillary Nanoliter Digestion 135
6.4 Conclusions 144
6.5 Literatures Cited 145

Chapter 7 147
Identification of *E. coli* Proteins and Protein Fragments by MALDI MS and Capillary LC MS/MS 147
7.1 Introduction 147
7.2 Experimental 148
7.3 Results and Discussion 150
7.4 Conclusions 164
7.5 Literatures Cited 168

Chapter 8 170
Conclusions and Future Work 170

List of Tables

Table 2.1 The m/z values of the common peaks observed for *B. thuringiensis* 10792 from two labs. 53

Table 2.2 The m/z values of the common peaks observed for *B. thuringiensis* 19267 from two labs. 54

Table 3.1 Optimized gradient separation conditions (shown as % solvent B. Solvent A: 0.05 %TFA in water; solvent B: 0.05% TFA in acetonitrile). 59

Table 3.2 $(M+H)^+$ from *E. coli* extract by online LC/ESI-MS. 62

Table 3.3 $(M+H)^+$ from *B. megaterium* extract by online LC/ESI-MS................62

Table 3.4 $(M+H)^+$ from *C. freundii* extract by online LC/ESI-MS. 62

Table 3.5 $(M+H)^+$ from *E. coli* crude mixture by direct MALDI. 65

Table 3.6 $(M+H)^+$ from *B. megaterium* crude extract by direct MALDI. 65

Table 3.7 $(M+H)^+$ from *C. freundii* crude extract by direct MALDI. 65

Table 3.8 $(M+H)^+$ from *E. coli* extract by LC/off-line MALDI. 66

Table 3.9 $(M+H)^+$ from *B. megaterium* extract by LC/off-line MALDI. 67

Table 3.10 $(M+H)^+$ from *C. freundii* extract by LC/off-line MALDI. 68

Table 3.11 $(M+H)^+$ of *E. coli* harvested at 36 h by direct MALDI (see text). 75

Table 4.1 Protein assay results for extraction lyopholized *E. coli* sample using different solvent extraction methods. 89

Table 4.2 Tryptic peptides matched protein FKBP-type peptidedyl-polyl cis-trans isomerase FKPA from gel band #3 [Figure 4.1A]............................93

Table 4.3 Proteins identified from gel band #2 of Figure 4.1A. 95

Table 4.4 Protein MH^+ detected from 0.1% TFA extracts (sonication or vortexing) using different MALDI matrices and sample preparation methods (see text), those marked with * match the results shown in the first column. 103

Table 5.1 Tryptic peptides matched 50S ribosomal protein L33 from fraction #124 116

Table 5.2 Chymotryptic peptides matched 50S ribosomal protein L33 from fraction #124. 117

Table 5.3 Summary of the identification of proteins from HPLC fractions by dual enzyme digestion. 117

Table 7.1 Proteins identified from Fraction #43 by LC MS/MS. 152

Table 7.2 Tryptic peptides matched protein UP04_ECOLI. 157

Table 7.3 Tryptic peptides matched YAHO_ECOLI and CSGA_ECOLI from Fraction #32. 163

Table 7.4 Proteins identified from HPLC fractions by LC MS/MS. 166

List of Figures

Figure 1.1 Schematic drawing of a typical prokaryotic cell. (Source: *http://www/bact.wisc.edu/Bact303*) 2

Figure 1.2 Schematic representation of gram-positive and gram-negative bacterial cell structure. (Source: *http://www.sp.uconn.edu/~terry/229sp00/lectures/cells2.html*)....2

Figure 1.3 Principle of matrix-assisted laser desorption ionization. 8

Figure 1.4 Schematic of a linear time of flight (TOF) mass spectrometer....10

Figure 1.5 Schematic of time-lag focusing. (A) Ions are desorbed from MALDI target with different initial kinetic energy. (B) Ions are expanded in field free region where high energy ions move further away from the repeller. (C) Ions reach the detector simultaneously due to the energy compensation. 11

Figure 1.6 Schematic of a reflectron (ion mirror) TOF mass spectrumeter. 13

Figure 1.7 Principle of the electrospray ionization process. 14

Figure 1.8 Schematic of a quadrupole mass spectrometer and the basic principle. 17

Figure 1.9 Schematic of a quadrupole ion trap mass spectrometer. 17

Figure 1.10 Nomenclature of peptide fragmentation pattern under low energy CID. ...21

Figure 2.1 MALDI spectra of *E. coli* 9637 obtained at the U of A lab by using different solvents for preparing the second-layer solution in a two-layer sample preparation method. The second-layer solution consists of saturated HCCA in (A) 33% acetonitrile/67% water and the sample solution (1:1) (all by volume), (B) 17% formic acid/33% isopropanol/50% water and the sample solution (1:1), and (C) 17% formic acid/33% methanol/50% water (FMW) and the sample solution (1:1). The sample solution is the bacterial extract using 0.1% TFA as the extraction solvent. 34

Figure 2.2 MALDI spectra of *B. thuringiensis* 19267 obtained at the ERDEC lab. (A) The bacterial extract was desalted by a membrane filter prior to the MALDI sample preparation, (B) no desalting step was used. A 50-mM ammonium bicarbonate solution was used for protein extraction. 37

Figure 2.3 MALDI spectra of *E. coli* 9637 obtained at the U of A lab under a controlled experimental condition to illustrate the effect of the type of extraction solvent on mass spectral pattern. (A) 0.1% TFA was used as the extraction solvent and (B) a mixture containing 17% formic acid/33% methanol/50% water (by volume) was used as the extraction solvent (see text for details on performing this extraction). The final solvent composition used for preparing the second-layer solution in MALDI was adjusted to be the same in both cases. 40

Figure 2.4 MALDI spectra of *E. coli* 9637 obtained at the U of A lab by using (A) methanol, (B) isopropanol, (C) water, (D) ammonium bicarbonate, and (E) Tris-HCl buffer as the extraction solvent. The second-layer solution in the two-layer sample preparation consisted of 17% formic acid/33% isopropanol/50% water and the bacteria extract. 41

Figure 2.5 MALDI spectra of *B. thuringiensis* 19267 obtained at the U of A lab by using ammonium bicarbonate solutions with different pH values: (A) pH=7.6, (B) pH=8.0, and (C) pH=8.5. 42

Figure 2.6 MALDI spectra of *B. thuringiensis* 10792 obtained at the U of A lab by using ammonium bicarbonate extraction solutions with different pH values: (A)pH=7.6, (B)pH=8.0, and (C) pH=8.5. 43

Figure 2.7 MALDI spectra of *E. coli* 11775 obtained by using different extraction methods: (A) using the solvent suspension method at the same condition as that of Figure 2.1B (collected at the U of A lab) and (B) using the enzyme-based extraction method (collected at the ERDEC lab). 45

Figure 2.8 MALDI mass spectra of *B. thuringiensis* 10792 obtained at U of A (Figure 4.8A) and ERDEC (Figure 4.8B). Common peaks are designated by solid circles. Ammonium bicarbonate solution was used for extraction and the two-layer method was used for sample/matrix preparation. 47

Figure 2.9 MALDI mass spectra of *B. thuringiensis* 19267 obtained at U of A (Figure 2.9A) and at ERDEC (Figure 2.9B). Common peaks are marked with solid circles. Ammonium bicarbonate solution was used for extraction and the two-layer method was used for sample/matrix preparation. In panel 2.9C is shown the mass spectrum obtained at ERDEC using the formic acid (F)/methanol (M)/water (W) extraction, wherein the major peaks are essentially replicates of those in panels 2.9A and 2.9B. To clarify, the approach used for Figure 2.9C involved sequential addition of the solution components in order F, M, W with vortexing between additions of components. If the FMW solution is prepared first, then added to the sample with vortexing, a significantly different intensity distribution is observed. 49

Figure 3.1 Total ion chromatogram of the *E. coli* extract separated by 2.1×150mm C_8 column with the static mixer of HP1100 removed. 63

Figure 3.2 MALDI spectra of bacterial proteins extracted by 0.1% TFA aqueous solution. (A) *E. coli*, (B) *B. megaterium*, (C) *C. freundii*. 64

Figure 3.3 MALDI spectrum of *E. coli* harvested at 36 h. The peaks with * matched the masses in LC off-line MALDI table (Table 3.8). 75

Figure 4.1 SDS PAGE of *E. coli* 9637 extracts prepared using different extraction solvents with vortexing: (A) 0.1% TFA, (B) 40 mM Tris-base (pH 9), and (C) 50 mM NH_4HCO_3 (pH 9). The proteins loaded into each lane were extracted from 0.2-0.5 mg starting lyophilized sample. Protein identification was carried out on the bands labeled with arrows (see text). 86

Figure 4.2 MALDI mass spectra of *E. coli* 9637 extracts prepared in different solvents with vortexing: (A) 0.1% TFA and (B) 40 mM Tris-base. 88

Figure 4.3 (A) MALDI mass spectrum of gel band #1 of Figure 4.1A after extraction. (B) MALDI mass spectrum of the in-gel digest of gel band #1. (C) ESI MS/MS spectrum of a tryptic peptide with MH^+ at 632.4 Da. 91

Figure 4.4 MS/MS spectrum of a doubly charged tryptic peptide with $[M+2H]^{2+}$ at 883.8 Da from DGAL_ECOLI. 94

Figure 4.5 SDS PAGE image of proteins extracted from different bacteria with 0.1% TFA: (A) *C. freundii*, (B) *A. hydrophilia*, and (C) *B. cereus*. The proteins loaded into each lane were extracted from 1 mg lyophilized bacterial samples. 96

Figure 4.6 SDS PAGE image of *E. coli* proteins using probe tip sonication in different extraction solvents: (A) 0.1% TFA and (B) 40 mM Tris-base. The proteins loaded into each lane were extracted from 0.2-0.5 mg lyophilized cells. ...97

Figure 4.7 MALDI analyis of *E. coli* extract prepared by probe tip sonication in 40 mM Tris base (A) spectrum of crude extract, (B) spectrum of extract after 10 kDa molecular mass cutoff. 98

Figure 4.8 MALDI spectra of *E . coli* extract in 0.1% TFA using probe tip sonication. (A) HCCA as matrix using two-layer MALDI sample preparation, (B) SA as matrix using two-layer MALDI sample preparation, (C) HABA as matrix using dried droplet sample preparation. 99

Figure 5.1 MALDI spectra of tryptic digestion on fraction #118 obtained by different second layer sample composition (A) tryptic digest mixed 1:1 (by volume) with the second layer matrix solution. (B) tryptic digest mixed 1:5 with the second layer matrix solution. 112

Figure 5.2 MALDI spectra of fraction #114 and its tryptic and chymotryptic digests. (A) molecular weight determination, (B) tryptic digestion, (C) chymotryptic digestion. 114

Figure 5.3 MALDI spectrum of fraction #124 shown a major component of MH^+ at 6255 Da. 116

Figure 5.4 LAP N-terminal digestion of intact protein from fraction #114. (A) MALDI spectrum showing the molecular ion and its methionine oxidized ions, (B) MALDI spectrum taken after 2 min LAP N-terminal digestion. 119

Figure 5.5 LAP N-terminal digestion on intact protein in fraction #118. (A) MALDI spectrum of fraction #118, (B) the spectrum taken after 2 min LAP digestion, (C) the spectrum taken after 5 min LAP digestion. 121

Figure 5.6 LAP N-terminal digestion on tryptic digest of fraction #118. (A) Tryptic digestion result showing the major peak with m/z at 1232.3; (B)The spectrum taken after 5 min LAP digestion on the tryptic digest of fraction #118. 122

Figure 6.1 Schematic drawing of in-capillary sample concentration and washing. 129

Figure 6.2 Schematic drawing of in-capillary reaction and microspot sample deposition. 130

Figure 6.3 Schematic comparison of macro- and micro-MALDI sample preparation. 131

Figure 6.4 MALDI mass spectra of in-capillary tryptic digests of a 4 μM cytochrome c solution in 40 mM NaCl and 20 mM NH_4HCO_3 buffer. (A) Direct deposition of digest mixture onto MALDI target without any washing step. (B) Simultaneous deposition of contaminated digest mixture and a matrix solution plug onto target. (C) Simultaneous deposition of digest mixture of washed protein and matrix solution onto target. 133

Figure 6.5 MALDI analysis of HPLC fraction #42 of *E. coli* extract. (A) Molecular weight determination, (B) MALDI peptide mass mapping. 136

Figure 6.6 MALDI mass spectra of in-capillary digests of cytochrome c. 137

Figure 6.7 Sections of MALDI mass spectra from in-capillary digests of fraction #52 containing DNA Binding Protein HU Alpha. (A) Only trypsin digest (B) Trypsin digest followed by LAP for 5 min. (C) Trypsin digest followed by LAP for 15 min. For each experiment a total volume of ~5 nL was concentrated in ~500 pL portions inside the capillary as described in the Experimental section. 140

Figure 6.8 MALDI mass spectra of in-capillary tryptic digests of fraction containing 50S ribosomal protein L31. 141

Figure 7.1 MALDI spectrum of fraction #43. 151

Figure 7.2 MS/MS spectra of tryptic peptides from EFTU_ECOLI.
(A) AFDQIDNKPEEK (position 46-57), (B) TTLTAAITTVLAK
(position26-38), (C) GITINTSHVEYDTPTR (position 60-75),
(D) TKPHVNVGTIGHVDHGK (position 9-25). 158

Figure 7.3 MALDI spectrum of fraction #32. 159

Figure 7.4 MS/MS spectra of two tryptic peptides from HNS_ECOLI. 161

Figure 7.5 MS/MS spectrum of an unexpected tryptic peptide from HNS_ECOLI.
....162

List of Abbreviations

HCCA	α-cyano-4-hydroxycinnamic acid
SA	sinapinic acid
HABA	2'-(4-hydrooxyphenylazo)benzoic acid
TFA	Trifluoroacetic acid
FMW	17% formic acid/33% methanol/50% water
LAP	leucine aminopeptidase
PepM	methionine aminopeptidase
DTT	dithiothreitol
Met	methionine
E. coli	*Escherichia coli*
B. megaterium	*Bacillus megaterium*
C. freundii	*Citrobacter freundii*
B. subtilis	*Bacillus subtilis*
B. thuringiensis	*Bacillus thuringiensis*
B. cereus	*Bacillus cereus*
A. hydrophilia	*Aeromonas hydrophilia*
UV	ultraviolet
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
HPLC	high performance liquid chromatography
MALDI	matrix-assisted laser desorption/ionization
ESI	electrospray ionization
MS	mass spectrometry
TOF	Time-of-Flight
FT-ICR	Fourier Transform Ion Cyclotron Resonance

Q	quadrupole
IT	quadrupole ion trap
CID	collision induced dissociation
PSD	post-source decay
MW	molecular weight
m/z	mass-to-charge ratio
nL	nanoliter (1 nL = 10^{-9} L)
µL	microliter (1 µL = 10^{-6} L)
µg	microgram (1µg = 10^{-6} g)
µm	micrometer (1 µm = 10^{-6} m)
TIC	total ion chromatogram
OM	outer membrane
BSA	bovine serum albumin

Chapter 1

Introduction to

Bacterial identification by Mass Spectrometric Techniques

Bacteria are unicellular microorganisms and the most studied prokaryotes. The fundamental difference between prokaryotic and eukaryotic cells is that prokaryotes do not have membrane-enclosed nucleus and other membranous organelles. In addition, prokaryotic cells usually have a more complex cell wall structure than eukaryotic cells. Prokaryotes are also generally smaller than eukaryotic cells; a typical bacterial cell is about 1 micrometer in diameter while most eukaryotic cells are from 10 to 100 micrometers in diameter. As a consequence, prokaryotic cells have a much higher cell surface to cytoplasm volume ratio compared to eukaryotic cells. Prokaryotes are the most abundant form of life on the planet, both in terms of biomass and total number of species. Because of their simplicity and our general high knowledge of their biological process, prokaryotic cells, especially bacterial cells, are commonly used as models for studying molecular biology, genetics, and physiology of all types of cells.

Figure 1.1 shows a typical prokaryotic cell structure. It has three major architectural regions: appendages (flagella and pili) outside the cell wall that are made of cell surface proteins; a cell envelope that consists of a capsule, cell wall and plasma membrane; and a cytoplasmic region that contains the cell genome (DNA), ribosomes and various sorts of inclusions. Bacterial cells exist in various shapes, such as rod (bacillus), spheres (coccus), spiral, filamentous and pleiomorphic shapes. Bacteria

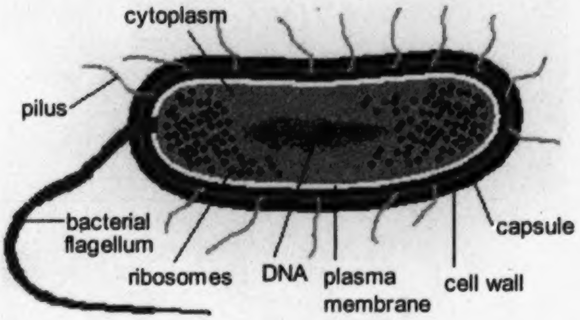
cytoplasm
pilus
bacterial flagellum
ribosomes
DNA
plasma membrane
cell wall
capsule

Figure 1.1 Schematic drawing of a typical prokaryotic cell. (Source: http://www/bact.wisc.edu/Bact303)

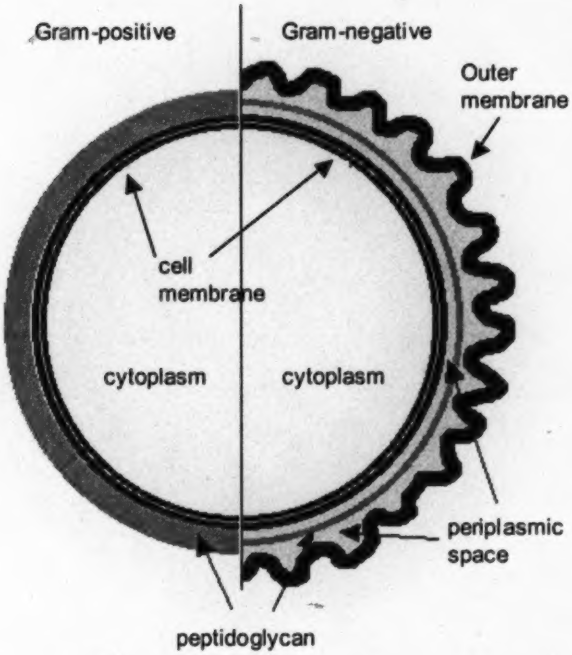
Gram-positive
Gram-negative
Outer membrane
cell membrane
cytoplasm
cytoplasm
periplasmic space
peptidoglycan

Figure 1.2 Schematic representation of Gram-positive and Gram-negative bacterial cell structure. (Source: http://www.sp.uconn.edu/~terry/229sp00/lectures/cells2.html)

may form pairs, chains, clusters or other groupings. Such formations are usually characteristic of a particular species. Bacteria are named according to the Linnaean system: genus + species. For example, *Escherichia* (genus) *coli* (species) (*E. coli*) and *Bacillus* (genus) *subtilis* (species) (*B. subtilis*).

Bacterial cells are classified by a Gram staining procedure as either gram-positive or gram-negative.[1] Gram-positive cells retain the color of crystal violet/iodine complex and remain purple after washing with alcohol; gram-negative cells do not retain the dye and they are colorless until counterstained with safranin dye and appear pink. The different reaction to the staining procedure is due to their different cell wall structures. As illustrated in Figure 1.2, gram-positive cells (such as *B. subtilis*) have a thick peptidoglycan layer which makes it much more rigid. Gram-negative cells (such as *E. coli*) have a much thinner peptidoglycan layer, but they have outer membrane (OM) made of lipid, proteins, and lipopolysaccharide. The OM is porous and small molecules can pass freely through it. Gram-negative bacterial cells also have a periplasmic space between the inner membrane and OM. It occupies up to 30% of the cell volume, in which many specialized proteins including enzymes and transport proteins are located. Gram-negative bacterial cells are generally more susceptible to cell rupture or lysis by mechanic forces than gram-positive cells. Some gram-positive bacteria form endospores when essential nutrients are depleted or when water is unavailable. Endospores possess thick walls or additional layers, which make them even more difficult to break by mechanical stress.

1.1 Bacterial identification by Mass Spectrometry Using Proteins as Biomarkers

The accurate and sensitive identification of bacterial pathogens is very critical in diagnosing diseases, identifying the nature of biological agents (for example, warfare or terrorist threats) and preventing potential biological and environmental hazards. Traditional microbiological techniques for the identification of bacteria are largely based on the determination of a diverse set of phenotypic characters, such as morphological, biochemical and physiological features, growth requirements, and chemical and serological properties. Given the great diversity of bacteria, it is often necessary to carry out a preliminary set of tests that lead to an appropriate subgroup. This process is both time-consuming and laborious.

An alternative approach, based on analytical measurement of chemical constituents (biomarkers) of bacterial cells, has become a potentially important tool in bacterial identification. These techniques are often referred to as chemotaxonomic methods.[2,3] Different cell components, such as lipids, carbohydrates, DNAs or proteins, have been used for chemotaxonomic analysis. The application of mass spectrometry (MS) to characterize bacteria and other microorganisms based on chemical biomarkers was first proposed in 1975.[4] Most of the early efforts used lipids, phospholipids, lipopolysaccharides, oligosaccharides, and oligonucleotides as biomarkers.[4-20] Pyrolysis products decomposed from bacterial biomolecules were also evaluated as possible biomarkers.[21]

The development of two modern ionization techniques, electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI), makes it practical to generate ions from large, nonvolatile analytes such as proteins and peptides without significant analyte decomposition or fragmentation. This offers a great opportunity to use proteins or peptides as biomarkers for chemotaxonomic characterization of bacteria. As potential biomarkers, proteins not only provide indirect, strain-specific genetic information about the bacteria, but they are also present in relatively high abundance in the cells. For example, proteins account for up to 70% of cell dried weight in *E. coli* and *Bacillus* spp.[22] Consequently, using proteins or peptides as biomarkers can potentially become more specific and sensitive. Moreover, the wealth of protein information for bacteria is being archived for searching in public databases, in which protein molecular mass as well as their amino acid sequence information can be easily accessed. MS analysis of proteins or subsets of the proteome (i.e., proteins expressed by the genome of a cell) should be a viable approach for unambiguous identification of bacteria.

Both MALDI and ESI MS have been used for bacterial identification based on their different protein components.[23-30] MALDI is more often used due to its tolerance to interference by salts and buffers and its ability to generate mainly singly charged species. However, ESI has gained more and more attention recently, because it can be easily interfaced with modern liquid phase separation techniques, and thereby dramatically reducing the ion suppression effect generally encountered when dealing with crude cell extracts. Bacterial proteins have been successfully analyzed by MS from either cell extracts [22, 23, 28, 31-39] or directly from whole cells.[24-26, 40-48]

Two possible routes have been used for bacterial identification based on the detection of protein biomarkers. One approach relies on producing mass spectral fingerprints of proteins and then comparing unknown species to the archived data or to the data from concurrent analysis of known bacteria.[22-25, 27, 28, 34, 40, 41] A crucial requirement for successful identification by this approach is mass spectral reproducibility. However, the spectra of such complex mixtures depend on a number of analytical and microbiological factors, such as bacterial growth environment, growth time, sample pretreatment, MS sample preparation as well as the MS instruments. Consequently, even for replicate experiments on the same bacterial culture, very different spectra may be obtained when the experimental conditions are not well controlled. An alternative approach involves detecting and cataloguing the masses of proteins expressed by bacteria. These proteins masses would then be searched against the public Internet based proteome database,[49] or databases created by mass spectrometric methods, or a combination of these two databases. Potential matches would be retrieved with statistically significant scores assigned to the possible candidates. Since a subset of proteins represented by their molecular masses are used in this approach, mass spectral reproducibility will not be a major concern. Different sets of protein biomarkers might be generated under different conditions. These sets should represent the genus, species and strain of individual bacteria. The successful identification by this approach, however, greatly depends on the availability and completeness of the bacterial protein information in the databases.

Unequivocal chemical identification of the commonly observable proteins from different bacteria is a fundamental issue for creating protein mass databases by MS. It will not only validate the methodology of using such databases, but also help in further optimizing sample preparation and MS analysis methods to improve the sensitive detection of those specific protein biomarkers. Protein identification is usually done by first separating proteins from the crude bacterial cell extracts, enzymatically digesting the purified proteins, and then using MALDI-TOF or LC/ESI-MS to analyze the digests. Protein identification can be achieved by either peptide mass mapping using the set of tryptic peptide masses, or by peptide sequence information obtained by MS/MS on specific tryptic peptides.

In the following sections, a brief overview will be given on different MS technologies and their applications for protein detection and identification.

1.2 Mass Spectrometry

Mass spectrometry is an analytical technique that measures the mass to charge ratio of individual ions. The analyte molecules are first converted into gas-phase ionic species, followed by separation of these ions in a mass analyzer, and then the ion current from the mass separated ions is detected by a suitable detector, and displayed in the form of a mass spectrum. The choice of a method for the ionization of a compound is contingent on the nature of the sample under investigation and the information required. As soft ionization techniques, MALDI and ESI are widely used for biopolymer analysis. These biopolymers are generally nonvolatile and thermally unstable.

1.2.1 MALDI Time-of-Flight (TOF) MS

MALDI was first reported in 1987 by two research groups independently.[50-53] The principle of this technique is shown schematically in Figure 1.3. A proper organic matrix is required for mixing with the analyte in a ratio of, generally, > 500:1. About 1 μL of the mixture is typically deposited onto a MALDI sample target. After drying, the analyte and matrix form co-crystals on the target, which is then inserted into a mass spectrometer. An ultraviolet (UV) laser beam of a short pulse of ~3 ns duration and a power of ~10^6 W/cm^2 is used to irradiate the sample target. A large amount of energy is absorbed efficiently by the chromophoric matrix molecules, which rapidly expand into the gas phase. The analyte molecules are desorbed together with the matrix molecules. Although there is no consensus on the ionization mechanism, it is widely accepted that, for proteins, ionization occurs via gas phase proton-transfer reactions between excited matrix molecules and analyte molecules.[54]

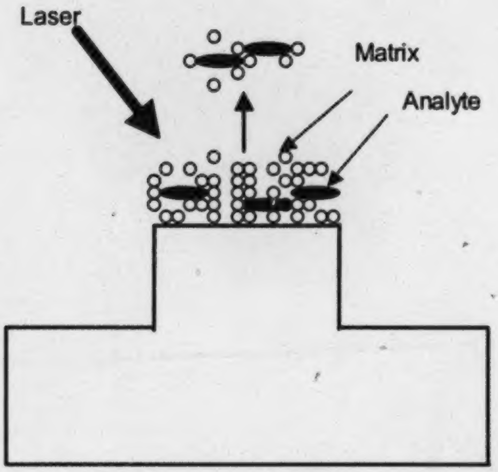

Figure 1.3 Principle of matrix-assisted laser desorption ionization (MALDI).

The matrix performs three important functions during the MALDI process: (1) it absorbs photon energy from the laser light; (2) it serves as a solvent for the analyte, isolating the biomolecules from each other in the solid crystal, thereby greatly reducing intermolecular interactions; (3) it provides photo-excited acidic or basic sites for ionization of sample molecules.

MALDI has several features that make it very suitable for biological applications. It produces primarily singly charged ions, allowing the analysis of very heterogeneous mixtures. Additionally, analyte is incorporated into matrix crystals upon the evaporation of the solvents whereas salts are excluded from the crystal and pushed into the rim of the sample layer. The crystallization process acts as an *in-situ* sample cleaning step, which partially accounts for the high tolerance of MALDI for various contaminants.

Many types of mass spectrometers are used with MALDI, including time-of-flight (TOF), Fourier Transform Ion Cyclotron Resonance (FT-ICR), quadrupole ion trap (IT), and magnetic sector. Among these, the TOF analyzer is most commonly coupled to MALDI because its pulsed ion detection mode is well matched with the pulsed ionization in the MALDI process. In addition, TOF's theoretically unlimited mass range, short duty cycle, high ion transmission, and multichannel detection features are also highly desirable for high sensitivity MALDI analysis.

Figure 1.4 shows the basic principle of a linear TOF analyzer. The ions generated in the ion source are accelerated by a voltage (V) of up to 30 kV before passing through a field free drift tube of 0.5-2 m in length (L). All the ions gain the same kinetic energy in the acceleration region. To a first approximation, the kinetic energy can be described as

$$z\,e\,V = \tfrac{1}{2}\,m\,v^2 \qquad (1.1)$$

where e is the unit of elementary charge, m is the mass of the ion, and z is the charge state and v is the linear velocity of the ion after acceleration.

In TOF, the mass to charge ratio (m/z) of an ion is determined by measuring its flight time (t), t = L / v. Therefore,

$$t = \left(\frac{m}{2 \cdot e \cdot V} \right)^{1/2} \cdot L \qquad (1.2)$$

Where L is the length of the field free region. Note that the starting point of the flight time is the point when the laser starts to irradiate the sample target. For simplicity, no initial kinetic energy was considered in this equation. The initial energy for ions generated by the MALDI process can vary significantly (see below).

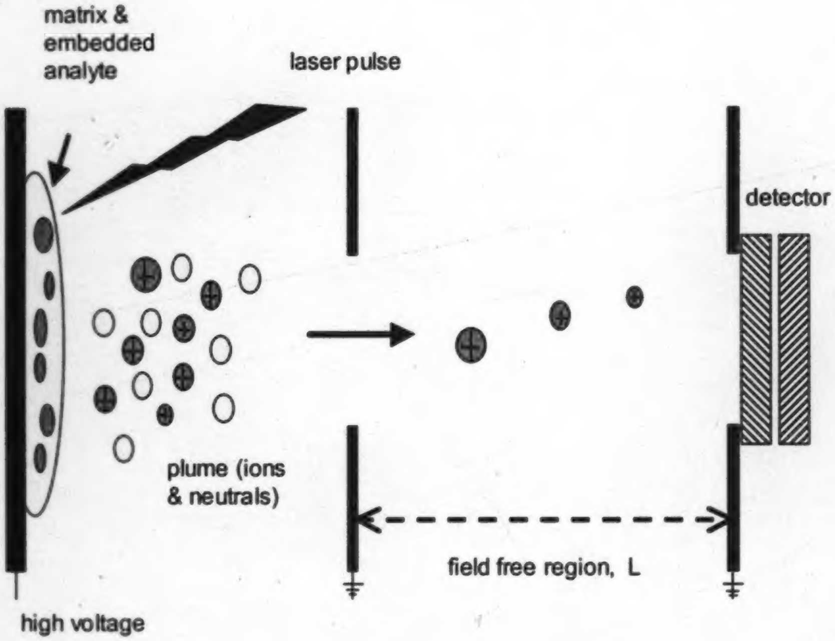
matrix &
embedded
analyte
laser pulse
detector
plume (ions
& neutrals)
field free region, L
high voltage

Figure 1.4 Schematic of a linear time of flight (TOF) mass spectrometer.

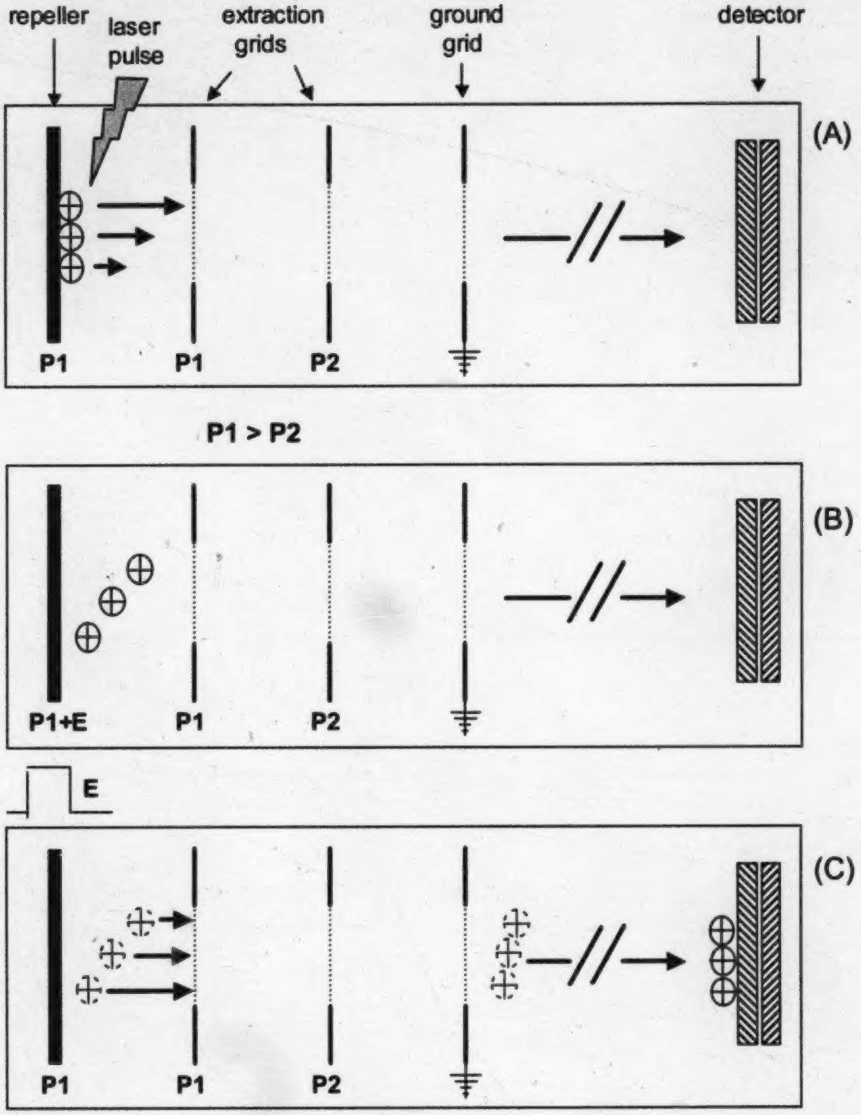
repeller
laser pulse
extraction grids
ground grid
detector
(A)
P1
P1
P2
P1 > P2
(B)
P1+E
P1
P2
E
(C)
P1
P1
P2

Figure 1.5 Schematic of time-lag focusing. (A) Ions are desorbed from MALDI target with different initial kinetic energy. (B) Ions are expanded in field free region where high energy ions move further away from the repeller. (C) Ions reach the detector simultaneously due to the energy compensation. P1 and P2 are the DC voltages applied on the repeller and extraction grids, E is the pulsed voltage applied on the repeller after a certain time delay.

The mass resolution ($R = m/\Delta m$) of a linear TOF mass spectrometer is usually very poor (generally less than 300). This is mainly due to the fact that ions formed by MALDI display a broad initial velocity or energy distribution, and they are always generated within a finite time and space. Peak broadening can be reduced by using a time-lag focusing TOF analyzer or a TOF analyzer equipped with a reflectron (ion mirror), or a combination of the two.

Figure 1.5 shows the principle of time-lag focusing in a linear TOF MS. In time-lag focusing, the repeller and the first extraction grid are held at the same potential for a certain time after the formation of ions (Figure 1.5A-B). During this "lag" period, ions of the same m/z but different initial axial velocities will move differently in the field free region between the repeller and the first extraction grid. The ions with high initial velocities move further from the repeller towards the first grid. After a certain time delay (typically hundreds of nanoseconds to several microseconds), an extraction pulse is applied to the repeller to extract the ions into the flight tube. The extraction pulse imparts more energy to the ions closer to the repeller such that the initially less-energetic ions that were closer to the repeller catch up with the initially more energetic ions at the detector (Figure 1.3C). Thereby peak broadening is greatly reduced.

A schematic illustration of a reflectron TOF mass analyzer is shown in Figure 1.6. Ions are decelerated in the reflectron and turn around at different locations in the reflecting electric potential gradient. The ions of higher kinetic energy penetrate deeper into the reflectron and take a longer period of time to return. With properly arranged geometry and voltages on the reflectron, the initial energy spreads are largely compensated, and the mass resolution is greatly improved.

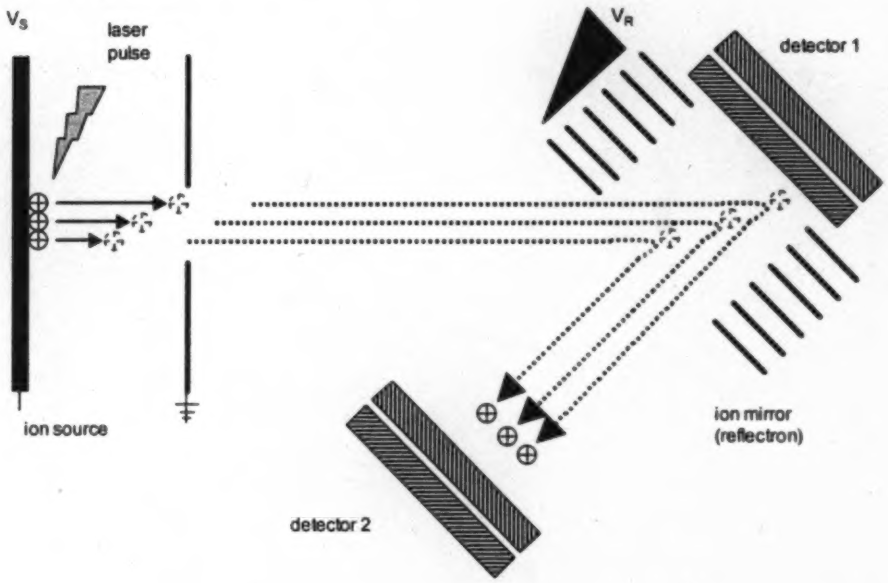

Figure 1.6 Schematic of a reflectron (ion mirror) TOF mass spectrumeter. Vs is the voltage applied on the repeller, V_R represents the voltages applied on the ion mirrors.

Besides the improved resolution, time-lag focusing and reflectron TOF MS also provide a significant improvement in mass measurement accuracy. Mass accuracy of better than 50 ppm can be routinely achieved for ions below 5000 Da. For complex mixture analysis over a wide mass range (5-70 kDa), mass accuracy of better than 500 ppm or 0.05% are usually obtained.

1.2.2 ESI-Quadrupole and ESI Ion Trap

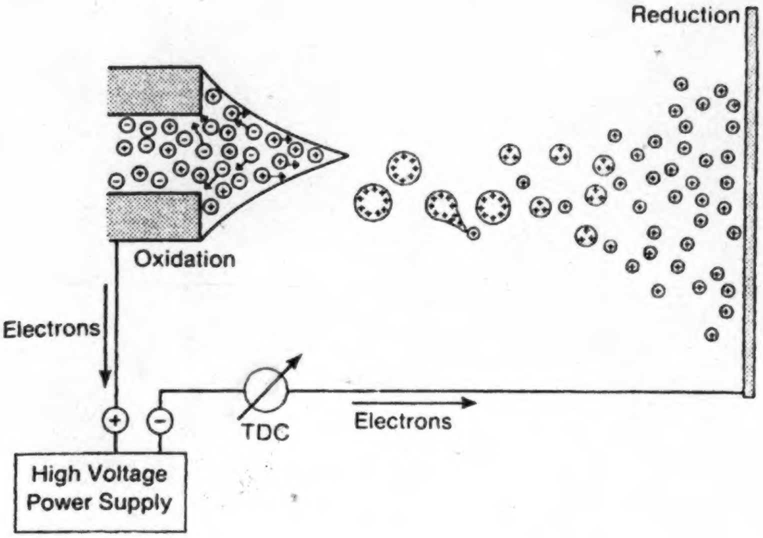

Figure 1.7 Principle of the electrospray ionization process.[58]

ESI is an atmospheric pressure ionization technique applicable to a wide range of compounds in solution. The concept of generating gas phase ions from electrically charged liquid droplets was first introduced by Dole in 1968.[55] Fenn and co-workers successfully coupled ESI to mass spectrometry in the mid-1980s.[56, 57] The ESI process was described by Kebarle et al. [58] as shown schematically in Figure 1.7. There are four major processes involved. (1) Charged droplets are formed at the ESI capillary tip by application of a high voltage relative to a counter electrode. (2) Charged droplets shrink due to evaporation in a sheath of dry nitrogen gas at moderate temperature. (3) As the droplets shrink, the coulombic repulsion forces become sufficient to overcome the surface tension forces and the charged droplets disintegrate into fine droplets. (4) The shrinking and disintegrating process repeats until it produces droplets so small that the electric field at the liquid surface becomes so high that the solute ions "escape" from the liquid phase to the gas phase (ion-evaporation), these ions are attracted into the mass spectrometer.

An improvement of electrospray, called ion spray, was introduced by Bruins and coworkers in 1987.[59] For ion spray, a concentrically applied nebulizer gas at the capillary tip is used to assist the formation of fine droplets. The substantial difference between ion spray and electrospray lies in the mechanism of droplet formation: whereas electrospray droplets are formed solely by the electric field on the capillary tip, droplets in ion spray are formed by a combination of the electric field and a jet of sheath gas. The main advantage of ion spray is that it can provide stable spray from 100% water to 100% organic modifier even for high flow rate analysis (i.e., 1 mL/min). This feature greatly facilitates the direct coupling of conventional liquid chromatography (LC) to ESI MS.

Since ESI produces ions continuously, it is often coupled to a scanning mass analyzer such as a quadrupole (Q) or a quadrupole ion trap (IT). The operation of both Q and IT is based on the motion of ions in an oscillating electric field.

A quadrupole mass analyzer consists of four cylindrical rods (Figure 1.8) to which the same absolute potential ($U+V\cos(\omega t)$) with different signs is applied. The potentials are set so that only ions within a small range of m/z have a stable trajectory and are transmitted to the detector. All the other ions collide with the rods and are pumped away. Ions of different m/z are scanned by simultaneously increasing the values of U and

V (U/V remains constant) such that ions are transmitted successively from low to high mass.

A quadrupole ion trap comprises two end cap electrodes and a ring electrode (Figure 1.9). A hyperbolic electric field is created inside the chamber by the potentials applied on the electrodes. An important aspect of the quadrupole ion trap is that ions can be accumulated, fragmented and mass analyzed in the same chamber. Ions generated externally are injected into the trap. With the proper setting of RF voltage on the ring electrode, ions within a certain mass range fall into the stable trajectory in the hyperbolic field and are trapped. Upon collision with the bath gas (e.g., helium), the ions tend to be focused into the center due to the loss of translational energy. These ions can be ejected out of the trap by using an ion selective instability operation mode or a resonance ejection mode and are then detected by an external detector.

The capability of performing tandem mass spectrometry makes ion trap very useful for peptide structure analysis. The ions of interest can be isolated by resonance ejection of all the other ions. The trapped ions are then fragmented by collision-induced dissociation (CID) and the product ions are scanned out of the trap and detected. MS^n (where n>2) experiments can also be performed using a quadrupole ion trap analyzer, and provide a wealth of structural information not obtainable by MS^2 scans.[60] The ability to perform MS^n experiments is very important for detailed characterization of proteins such as those with post-translational modifications.

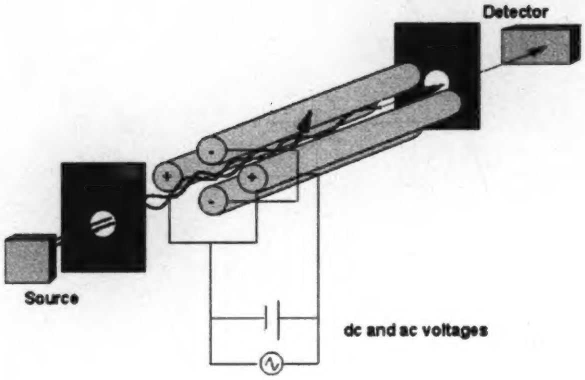
Detector
Source
dc and ac voltages

Figure 1.8 Schematic of a quadrupole mass spectrometer and the basic principle.

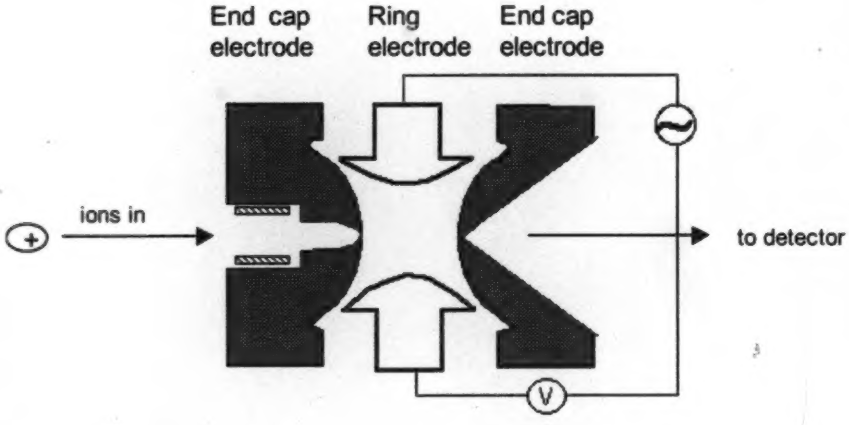
End cap electrode
Ring electrode
End cap electrode
ions in
to detector
V

Figure 1.9 Schematic of a quadrupole ion trap mass spectrometer.

1.3 Protein Identification by Mass Spectrometry

MS has become a routine technique for protein identification due to its speed and high sensitivity. Two levels of information are usually obtained by MS for protein identification. The first level is based on the masses of proteins and their proteolytic peptides – known as peptide mass fingerprinting or peptide mass mapping.[61-63] In this method, peptide masses generated from proteins digested by site-specific enzymes are used to search the proteome database, and the possible protein candidates are retrieved. The molecular masses of the intact proteins are often used to further constrain the searching process to obtain unequivocal results. The second level of information, peptide fragmentation pattern, is obtained by CID using tandem MS. The availability of commercial software, such as Sequest (Finnigan) and Biotool (Bruker), greatly facilitates the MS/MS spectral interpretation and database searching processes.

1.3.1 Peptide Mass Mapping

The sets of peptides generated by specific enzymatic or chemical cleavage of the intact proteins can be used as unique fingerprints to identify proteins. The peptide map can be produced by either MALDI or ESI MS. Compared to ESI, MALDI shows more tolerance to samples with buffers and salts and is more sensitive. In addition, MALDI almost exclusively produces singly charged ions for low mass peptides, which makes it more suitable for direct mixture analysis. Therefore, MALDI is usually the preferred method for peptide mass mapping. The peptide mass fingerprint obtained is then compared to the theoretical digest of the proteins in the proteome database, and the protein generating the most similar pattern is retrieved as a candidate. Several search engines in the Internet can be used for this purpose. These include MOWSE (http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse), MS-Fit (http://prospector.ucsf.edu/)

PeptideSearch (http://www.expasy.ch/tools/peptident.html) and ProFound (http://prowl.rockefeller.edu/cgi-bin/ProFound).

The accuracy of mass measurement of the peptides is very important for an unambiguous identification, especially when dealing with large size proteome databases. The implementation of time-lag focusing as well as reflectron TOF allows high resolution mass analysis in TOF. As a result, mass accuracy has been significantly improved. A modern MALDI-TOF instrument can routinely provide mass accuracy of better than 50 ppm with external mass calibration and better than 20 ppm with internal calibration. The increased mass accuracy dramatically reduces the number of matching peptides, thereby reducing the chance of a false positive identification.

Although peptide mass mapping is usually the first experiment to be carried out for protein identification, ambiguity can still exist. This is especially true with protein mixtures or when dealing with a very small amount of protein. In such cases, it is advantageous to use supplementary information, such as short peptide sequence tags or peptide fragmentation patterns, to constrain the database search and thus improve the confidence of identification.

1.3.2 Short Protein or Peptide Sequence Tag

A wide variety of exopeptidases are available, which can be used to gradually cleave the amino acids from either the carboxy-terminal or amino-terminal of a protein or peptide.[64-66] These enzymes can be used, following the endoproteolytic digestion of a protein, to create C-terminal or N-terminal peptide sequence tags. Since only limited amino acid residues (1-4) can be cleaved from the peptides, the spectra are usually very easy to interpret manually even without pre-fractionation of the endoenzymatic peptides.

A partial sequence can be reconstructed by comparing the spectra obtained with and without exopeptidase at different periods of exoproteolytic digestion. The short sequence tags can be used to improve the confidence for protein identification.

1.3.3 Peptide Fragment Ion Fingerprinting by MS/MS

Instead of relying on a set of proteolytic peptide masses for protein identification, an alternative and, sometimes, complementary approach is to use peptide fragment ion fingerprints generated by tandem mass spectrometry.[67-71] In tandem MS, the proteolytic peptide of interest (i.e., the parent ion) can be selected by the first stage mass scan, followed by CID in a collision cell where the ions undergo collisions with an inert gas (e.g., helium or argon). The resulting fragment ions are analyzed in the second stage mass scan. Alternatively, the fragment ion spectra can be generated by post-source decay (PSD) of peptide ions in a reflectron MALDI-TOF instrument. It can also be obtained by CID using ESI MS/MS in a triple quadrupole or an ion trap mass spectrometer. Compared to ESI MS/MS, which can be used to generate CID fragment ion spectra for most proteolytic peptides, MALDI PSD fragment ion spectra can only be obtained for a selective number of peptides. There is no apparent correlation between the peptide structure and the probability of undergoing PSD. Thus, PSD is not routinely used as a tool for peptide fragment ion fingerprinting. However, one advantage of using PSD is that it can be quite convenient and sensitive, if peptides do fragment. After MALDI analysis of peptide masses and if peptide mass mapping does not result in unambiguous identification of the protein, the same sample can be subjected to PSD in the same instrument to produce fragment ion spectra.

Algorithms for searching sequence databases using uninterpreted MS/MS spectra have been developed. The most widely used software is Sequest. The program searches for all the peptides in database which have the same mass as the parent ion, and then matches the predicted MS/MS spectrum with the experimentally determined one. Finally, it carries out a cross-correlation analysis of the best scoring peptide to determine the best match. Some other programs, such as MS-Tag and Fragfit available from the web, can also be used for this purpose, albeit not in an automated manner in most commercial tandem MS instruments.

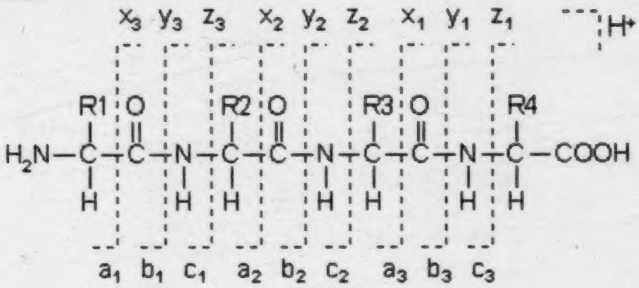
x3 y3 z3 x2 y2 z2 x1 y1 z1
H+
R1 O R2 O R3 O R4
H2N—C—C—N—C—C—N—C—C—N—C—COOH
H H H H H H H
a1 b1 c1 a2 b2 c2 a3 b3 c3

Figure 1.10 Nomenclature of peptide fragmentation pattern under low energy CID.

Two fragmentation-energy regimes (high or low energy CID) that differ in the amount of energy used for the fragmentation are employed for tandem mass spectrometry. Low energy CID is widely used by most instruments (triple quadrupole, ion trap and quadrupole-TOF) in protein MS laboratories. Peptide fragmentation patterns depend on the experimental parameters for CID and peptide sequence. Under CID, peptide ions fragment mostly along the peptide backbone. Figure 1.10 describes the nomenclature of ion series generated by CID.[72,73] The x, y, z series refer to the fragment ions in which the charge retains at the C-terminus, whereas for a, b, c series, the charge retains on the N-terminus. For low energy CID, b and y ion series are most often observed, with less frequent a and z ions. The ions resulting from the loss of water, ammonia or carbon monoxide from the sequence fragment ions are also often present.

If peptide mass mapping and peptide fragment ion fingerprinting fail to identify the protein, *de novo* sequencing of peptides can be carried out by using tandem MS. This usually involves manual interpretation of the fragment ion spectra to determine the amino acid sequence. This is a time consuming process and automation of the spectral interpretation is a subject of intense research at present. Once the sequence of one or more peptides from an unknown protein is obtained, database searching based on the genome sequence or cross species proteome sequence can be carried out to identify the gene sequence and the expressed protein.

1.4 Brief Summary of the Thesis

In this thesis, I first investigated several experimental factors related to the mass spectral reproducibility in direct MALDI analysis of proteins and peptides from crude bacterial cell extracts. Although experimental method related variations can be well controlled, variations associated with the biology of the bacteria, such as cell growth conditions, are much more difficult to control. This poses a major problem for bacterial identification based on the mass spectral pattern. A method of identifying bacteria based on searching a set of protein masses obtained by MALDI or ESI against a protein mass database of different bacteria created by MS methods was proposed. To achieve a confident bacterial identification, it is important to generate a large set of protein masses by MALDI analysis of cell extracts. Several factors affecting protein extraction efficiency were examined by gel electrophoresis. Their effects on the resulting MALDI spectra were also studied. Finally, efforts on bacterial protein identification are presented. Positive protein identification greatly validates our proposed methodology for bacterial identification.

1.5 Literatures Cited

1. Tortora, G. J.; Funke, B. R.; Case, C. L. *Microbiology An Introduction*, The Benjamin/Cummings Publishing Company, **1989**, pp 65-66.

2. Goodfellow, M.; Minnikin, D. D.; *Chemical Methods in Bacterial Systematics*. Academic Press, London, **1985.**

3. Gottschalk, G. *Methods in Microbiology*, Academic Press, London, **1985**, Vol 18.

4. Anhalt, J. P.; Fenselau, C. *Anal. Chem.* **1975**, 47, 219.

5. Heller, D. N.; Fenselau, C.; Cotter, R. J.; Demirev, P.; Olthoff, J. K.; Honovich, J.; Uy, M.; Tanaky, T.; Kishimoto, K. *Biochem. Biophys. Res. Commun.* **1987**, 142, 194.

6. Ho, B. C.; Fenselau, C.; Hansen, G.; Larsen, J.; Daniel, A. *Clin. Chem.* **1983**, 29, 1349.

7. Fenselau, C.; Cotter, R. J. *Chem. Rev.* **1987**, 87, 501.

8. Heller, D. N.; Cotter, R. J.; Fenselau, C.; Uy, O. M. *Anal. Chem.* **1987**, 59, 2806.

9. Black, G. E.; Fox, A.; Fox, K.; Snyder, A. P.; Smith, P. B. W. *Anal. Chem.* **1994**, 6, 4171.

10. Cole, M. J.; Enke, C. G. *Anal. Chem.* **1991**, 63, 1032.

11. Smith, P. B. W.; Snyder, A. P.; Harden, C. S. *Anal. Chem.* **1995**, 67, 1824.

12. Wunschel, D. S.; Fox, K. F.; Fox, A.; Bruce, J. E.; Muddiman, D. C.; Smith. R. D. *Rapid Commun. Mass Spectrom.* **1996**, 10, 29.

13. Hurst, G. B.; Doktycz, M. J.; Vass, A. A.; Buchanan, M. V. *Rapid Commun. Mass Spectrom.* **1996**, 10, 377.

14. Doktycz, M. J.; Hurst, G. B.; Habibi-Goudarzi, S.; Mcluckey, S. A.; Tang, K.; Chen, C. H.; Uziel, M.; Jacobson, K. B.; Woychik, R. P.; Buchanan, M. V. *Anal. Biochem.* **1995**, 230, 205.

15. Schleiffer, K. H.; Ludwig, W.; Amann, R. *Handbook of New Bacterial Systematics*, Academic Press, San Diego, CA, **1993**, 463.

16. Saiki, R. K.; Scharf, S.; Falcona, F.; Mullis, K. D.; Horn, G. T.; Erlich, H. A.; Amheim, N. A. *Science*, **1985**, 230, 1350.

17. Saiki, R. K.; Gelfund, P. H.; Stoffel, S.; Scharf, S. J.; Higuchi, R.; Horn, G. T.; Mullis, K. B.; Erlich, H. A. *Science*, **1988**, 239, 487.

18. Starnach, M. N.; Falkow, S.; Tompkins, L. S. *J. Clin. Micro*biol **1989**, 27, 1257.

19. Hance, A. J.; Grandchamp, B.; Levy-Frebault, V.; Lecosier, D.; Rauzier, J.; Bocart, D.; Gicquel, B. *Molec. Microbiol.* **1989**, 3, 843.

20. Fox, A.; Rogers, J. C.; Fox, K. F.; Schnitzer, G.; Morgan, S. L.; Brown, A.; Aono, R. J. *Clin. Microbiol.* **1990**, 31, 546.

21. Deluca, S.; Sarver,W. E.; Harrington, P. D.; Voorhees, K. J. *Anal. Chem.* **1990**, 62, 1465.

22. Krishnamurthy, T.; Ross, P.; Rajamani, U.; *Rapid Commun. Mass Spectrom.* **1996**, 10, 883.

23. Cain, T. C.; Lubman, D. M.; Weber, W. J. Jr. *Rapid Commun. Mass Spectrom.* **1994**, 8, 1026.

24. Holland, R. D.; Wilkes, J. G.; Rafii, F.; Sutherland, J. B.; Persons, C. E.; Voorhees, K. J.; Lay Jr., J. O. *Rapid Commun. Mass Spectrom*. **1996**, 10, 1227.

25. Claydon, M. A.; Davey, S. N.; Edward-Jones, V.; Gordon, D. B.; *Nature Biotechnol*. **1996**, 14, 1584.

26. Krishnamurthy, T.; Ross, P. L.; *Rapid Commun. Mass Spectrom*. **1996**, 10, 1992.

27. Arnold, R. J.; Reilly, J. P. *Rapid Commun. Mass Spectrom*. **1998**, 12, 630.

28. Wang, Z.; Russon, L.; Li, L.; Roser, D. C.; Long, S. R. *Rapid Commun. Mass Spectrom*. **1998**, 12, 456.

29. Krishnamurthy, T.; Davis, M. T.; Stahl, D. C.; Lee, T. D. *Rapid Commun. Mass Spectrom*. **1999**, 13, 39.

30. Dunlop, K. Y.; Li, L.; *J. Chromatogra A*, **2000**, 925, 123.

31. Chong, B. E.; Wall, D. B.; Lubman, D. M.; Flynn, S. J. *Rapid Commun. Mass Spectrom*. **1997**, 11, 1900.

32. Van Adrichem, J. H. M.; Bornsmen, K. O.; Conzelmann, H.; Gass, M. A. S.; Eppenberger, H.; Kresbach, G. M.; Ehrat, M.; Leist, C. H. *Anal. Chem*. **1998**, 70, 923.

33. Easterling, M. L.; Colangelo, C. M.; Scott, R. A.; Ameter, I. J. *Anal. Chem*. **1998**, 71, 3226.

34. Dai, Y.; Li, L.; Roser, D. C.; Long, S. R. *Rapid Commun. Mass Spectrom*. **1999**, 13, 73.

35. Holland, R. D.; Duffy, C. R.; Rafii, F.; Sutherland, J. B.; Heinze, T. M.; Holder, C. L.; Voorhees, K. J.; Lay Jr., J. O. *Anal. Chem*, **1999**, 71, 3226.

36. Arnold, R. J.; Reilly, J. P. *Anal. Biochem*. **1999**, 269, 105.

37. Birmingham, J.; Demirev, P.; Ho, Y.; Thomas, J.; Bryden, W.; Fenselau. C. *Rapid Commun. Mass Spectrom*. **1999**, 13, 604.

38. Wall, D. B.; Lubman, D. M.; Flynn, S. J. *Anal. Chem*. **1999**, 71, 3894.

39. Domin, M. A.; Welhan, K. J.; Ashton, D. S. *Rapid Commun. Mass Spectrom*. **1999**, 13, 222.

40. Welham, K. J.; Domin, M. A.; Scannel, D. E.; Cohen, E.; Ashton, D. S. *Rapid Commun. Mass Spectrom*. **1998**, 12, 176.

41. Haag, A. M.; Taylor, S. N.; Johnston, K. H.; Cole, R. B.; *J. Mass Spectrom*. **1998**, 33, 750.

42. Arnold, R. J.; Karty, J. A.; Ellington, A. D.; Reily, J. P. *Anal. Chem*. **1999**, 71, 1990.

43. Saenz, A. J.; Pertersen, C. E.; Valentine, N. B.; Gantt, S. L.; Jarman, K. H.; Kingsley, M. T.; Wahl, K. L *Rapid Commun. Mass Spectrom*. **1999**, 13, 1580.

44. Lynn, E. C.; Chung, M.; Tsai, W.; Han, C. *Rapid Commun. Mass Spectrom*. **1999**, 13, 2022.

45. Leenders, F.; Stein, T. H.; Kablitz, B.; Franke, B. P.; Vater, J. *Rapid Commun. Mass Spectrom*. **1999**, 13, 943.

46. Winkler, M. A.; Uher, J.; Cepa, S. *Anal. Chem*. **1999**, 71, 3416.

47. Evason, D. J.; Claydon, M. A.; Gordon, D. B. *Rapid Commun. Mass spectrom*. **2000**, 14, 669.

48. Madonna, A. J.; Basile, F.; Ferrer, I.; Meetani, M. A.; Rees, J. C.; Voorhees, K. J. *Rapid Commun. Mass Spectrom*. **2000**, 14, 2220.

49. Demirev, P. A.; Ho, Y.; Ryzhov, V.; Fenselau, C. *Anal. Chem*. **1999**, 71, 2732.

50. Karas, M.; Bachmann, D.; Bahr, Y.; Hillenkamp, F. *Int. J. Mass Spectrom. Ion Processes*, **1987**, 78, 53.

51. Karas, M.; Hillenkamp, F. *Anal. Chem.* **1988**, 60, 2299.

52. Tanaka, K.; Ido, Y.; Akita, S. In *Proceedings of the Second Japan-China joint Symposium on Mass Spectrometry*, Matsuda, H.; Liang, X. T. Eds, Bando Press, Osaka, Japan, **1987**, 185.

53. Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T. *Rapid Commun. Mass Spectrom.* **1988**, 2, 151.

54. Bõkelmann, V.; Spengler, B.; Kaufmann, R. *Eur. Mass Spectrom.* **1995**, 1, 81.

55. Dole, M. *J. Chem. Phys*, **1968**, 49, 2240.

56. Yamashita, M.; Fenn, J. B. *J. Phys. Chem.* **1984**, 88, 4671.

57. Whitehouse, C. M.; Dreyer R. N.; Yamashita, M.; B. Fenn, J. B. *Anal. Chem.* **1985**, 57, 675.

58. Kebarle, P.; Tang. L. *Anal. Chem.* **1993**, 65, 972A.

59. Bruins, A. P.; Covey, T. R.; Henion, D. *Anal. Chem.* **1987**, 59, 2642.

60. Strife, R. J.; Schwartz, J.; Bier, M.; Zhou, J. In *Proc. of the 43rd Conference on Mass Spectrometry and Allied Topics*, Atlanta, GA, May 21-26, 1995, pp.160.

61. Gibson, B. W.; Biemann, K. *Proc. Natl. Acad. Sci.* **1984**, 81, 1956.

62. Morris, H. R.; Panico, M.; Etienne, T.; Tippins, J.; Girgis, S. I. Macintyre, I. *Nature*, **1984**, 308, 746.

63. Greer, F. M.; Morris, H. R.; Forstrom, J.; Lyons, D. *Biomed. Environ. Mass Spectrom.* **1988**, 16, 191.

64. Doucette, A.; Li, L. *Joint-Issues of Proteomics and European Journal of Mass Spectrometry*, **2001**, in press.

65. Staudenmann, W.; Hatt, P. D.; Hoving, S.; Lehmann, A.; Kertesz, M.; James, P. *Electrophoresis* **1998,** *19*, 901.

66. Korostensky, C.; Staudenmann, W.; Dainese, P.; Gonnet, G.; James, P. *Electrophoresis* **1998,** *19*, 1933.

67. Mclafferty, F. W., Ed.; *Tandem Mass Spectrometry*, Wiley-Interscience: New York, 1983.

68. Busch, K. L.; Gilsh, G. L.; Mcluckey, S. A. *Mass Spectrometry. Techniques and Applications of Tandem Mass Spectrometry*; VCH: New York, 1988.

69. Biemann, K.; Scoble, H. A. *Science*, **1987**, 237, 992.

70. Biemann, K. *Biomed. Environ, Mass Spectrom*. **1988**, 16, 99.

71. Yates, J. R., II; Specicher, S.; Griffin, P. R.; Hunkapiller, T. *Anal. Biochem*. **1993**, 214, 397.

72. Roepstorff, P.; Fohlman, J. *Biomed. Mass Spectrom*. **1984**, 11, 601.

73. Biemann, K. *Methods Enzymol*. **1990**, 193, 886.

Chapter 2

Investigation of Spectral Reproducibility in Direct Analysis of Bacterial Proteins by MALDI-TOF MS[a]

2.1 Introduction

The potential of using MALDI time-of-flight mass spectrometry (TOF MS) for rapid identification of bacteria based on mass spectral patterns derived either from protein extracts [1,2] or from whole cells [3,4] has been widely studied. One or more peaks detected in the mass spectrum, which appeared to be conserved for a certain type of bacteria, were proposed as biomarkers. The protein biomarkers should be readily analyzed and less prone to interference from background species and molecules present in real world sample. Within each of these studies, the mass spectral patterns associated with different bacteria were different, providing evidence that protein analysis by MS can serve as a basis for bacterial discrimination. Identification of specific mass peaks as genus-, species-, even strain-, specific "biomarkers" has been attempted,[2,4] even though the database among which these identifications were made was limited in scope.

In order for MALDI and any other mass spectrometric methods to be used for bacterial identification, several important issues need to be addressed. Most important is the level of mass spectral reproducibility when different investigators use nominally the same protocols to study the same bacterial samples. An essential related concern is the extent to which experimental conditions affect the mass spectra. Only once these issues are addressed and understood, will it become possible to systematically seek and identify those bacterial peptide/protein masses which have validity as "biomarkers" for the discrimination of bacteria. In this chapter, we attempt to address the spectral reproducibility issue as well as to examine the effects of sample/matrix preparation and protein extraction method on mass spectral patterns.

---

[a] A form of this chapter is published as: Z. Wang, L. Russon, L. Li, D. C. Roser, S. R. Long., "*Investigation of Spectral Reproducibility in Direct Analysis of Bacteria Proteins by Matrix-assisted Laser Desorption/Ionization Time-of-flight Mass Spectrometry*" *Rapid Commun. Mass Spectrom.* **1998**, 12, 456-464.

2.2 Experimental

2.2.1 Chemicals and Material

The samples used here were two strains each of *Eschericia coli* (*E. coli*) (ATCC 9637 and 11175) and *Bacillus thuringiensis* (*B. thuringiensis)* (ATCC 10792 and 19267). The bacteria were grown overnight (18-24 h) in Nutrient Broth at 25 °C with shaking. Cells were harvested, washed with several volumes of sterile water, lyophilized to dryness, and stored at 0 C until use. Bacterial samples were grown at the Edgewood RDE Center (ERDEC), Aberdeen Proving Ground, MD, USA. Trifluoroacetic acid (TFA), isopropanol, α-cyano-4-hydroxycinnamic acid (HCCA), sinapinic acid (SA), horse cytochrome c, bradykinin, bovine insulin chain B, Tris, and ammonium bicarbonate were from Sigma-Aldrich-Fluka (Oakville, Ontario, Canada).

2.2.2 Extraction of Bacterial Proteins

The suspension solvents tested for dissolving proteins from the cells include water, 0.1% TFA in water, isopropanol, 50 mM ammonium bicarbonate, 10 mM Tris-HCl buffer, and a solvent mixture consisting of 10% formic acid/45% methanol/45% water (by volume). About 1 to 1.5 mg of lyophilized bacteria was suspended in 250 μl solvent (100 μL for the two basic solvents). The cell suspension was vortexed for about

2 min and then centrifuged at 14000rpm for 5 min. The supernatant solution was taken for MALDI analysis.

2.2.3 MALDI Analysis

The two-layer method was used for matrix/sample preparation.[5] HCCA and SA were initially examined as the matrices for bacterial protein analysis. It was found that HCCA provided better sensitivity over a wider mass range. Therefore, HCCA was used throughout this study. In the two-layer method, the first layer was formed by placing 1 μl of 100 mM HCCA in 99% acetone and 1% water (v/v) on the MALDI probe tip and allowing it to dry in air. For the second-layer solution, a saturated solution of HCCA was prepared in a solvent mixture. A variety of solvent mixtures were used. Their compositions are described in the Results and Discussion section. The saturated solution was routinely mixed with the sample solutions in a volume ratio of 4:1 in the cases of using methanol, isopropanol, ammonium bicarbonate, and Tris-HCl buffer as the extraction solvent, or 1:1 in the cases involving other extraction solvents. One microliter of this solution was placed on top of the first layer and allowed to dry. The probe tip was then dipped in pure water for approximately 10 s and the excess water was shaken off. Variations from these procedures are detailed as appropriate in the Results and Discussion section.

In the University of Alberta experiments, MALDI mass spectra were recorded in a time-lag focusing linear time-of-flight mass spectrometer with a 1-m flight tube.[6] A nitrogen laser (337 nm) was used for desorption/ionization. Mass spectra shown in this work represent the sum of 20 to 100 individual spectra. Spectra were mass calibrated externally with the use of bradykinin, bovine insulin chain B and horse cytochrome c as the calibrants. In the ERDEC experiments, spectra were recorded on a Vestec MALDI-TOF in linear mode using a nitrogen laser (337 nm) for desorption/ionization, with typically 40 to 100 individual scans accumulated per spectrum. The spectra were externally calibrated using a matrix dimer ion peak and cytochrome c, or calibration mixture 2 (angiotensin I, ACTH [1-17 clip], ACTH [18-39 clip], ACTH [7-38 clip], and bovine insulin) from the Sequazyme Peptide Mass Standards Kit (PerSeptive Biosystems).

2.3 Results and Discussion

Many experimental factors can influence the final appearance of the bacterial protein profile obtained by MALDI. In order to get reproducible spectra, these factors must be carefully identified and controlled. This is done through systematic investigation using different experimental conditions, with an emphasis on inter-laboratory comparison. Comparison of experimental results obtained from two laboratories has greatly facilitated the identification of sources of spectral discrepancy that may arise in analyzing the same samples. In both labs, sample preparation was based on the two-layer matrix/analyte preparation method. This method was found to be particularly useful for detecting mixtures of peptides and proteins covering a broad mass range.[7,8] For the majority of this work, a simple solvent suspension method was used for extracting peptides and proteins from the bacteria samples as detailed in Section 2.2.2.

2.3.1 Effect of the Sample Solvent

In the two-layer method, the first layer is the matrix crystals and the second layer is formed from a mixture of the analyte and matrix solution. Both the chemical composition and the type of solvent used to prepare the second layer solution can have a significant effect on the mass spectral pattern. Figure 2.1 shows the MALDI mass spectra of *E. coli* 9637 obtained using different solvents for preparing the second layer. In all three spectra, the cell suspension was prepared in 0.1% TFA and the first matrix layer on the probe tip was prepared using HCCA. The second-layer solution used for generating Figure 2.1A consisted of saturated HCCA in 33% acetonitrile/67% water mixed with the sample solution in a 1:1 ratio (all by volume). For Figure 2.1B, the second-layer solution consisted of saturated HCCA in 17% formic acid/33% isopropanol/50% water and the sample solution (1:1) (all by volume). For Figure 2.1C, the solution consisted of saturated HCCA in 17% formic acid/33% methanol/50% water (FMW) and the sample solution (1:1) (all by volume). The m/z values of major peaks in each spectrum are shown in the figures. Many common peaks are detected in the three spectra. However, the relative intensities of these peaks are different. Similar observations were obtained from the other bacterial samples examined. The salient feature is that a number of common peaks, albeit with different relative intensities, are observed under different sample/matrix preparation conditions. It appears that, with the two-layer method, variation of the solvent conditions for matrix/sample preparation has a much greater effect on the relative intensities of the peaks than on the actual number of proteins detected.

The observation of differences in overall detection sensitivity and relative peak intensity under different sample preparation conditions is not surprising.[5,9-15] The analyte incorporation and distribution in the matrix crystals can be affected by sample preparation.[5] The results shown in Figure 2.1 illustrate that it is important to control the solvent conditions used for preparing the samples in order to generate reproducible

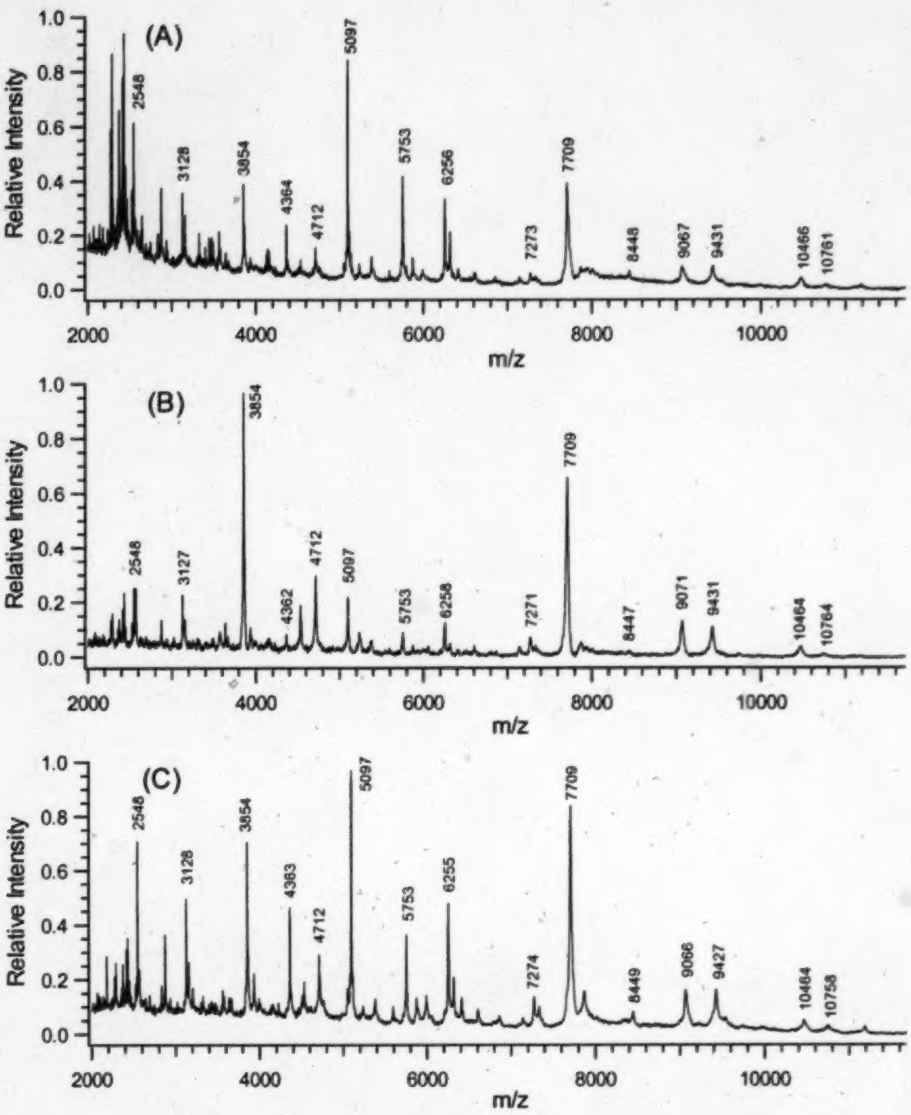

Figure 2.1 MALDI spectra of *E. coli* 9637 obtained at the U of A lab by using different solvents for preparing the second-layer solution in a two-layer sample preparation method. The second-layer solution consists of saturated HCCA in (A) 33% acetonitrile/67% water and the sample solution (1:1) (all by volume), (B) 17% formic acid/33% isopropanol/50% water and the sample solution (1:1), and (C) 17% formic acid/33% methanol/50% water (FMW) and the sample solution (1:1). The sample solution is the bacterial extract using 0.1% TFA as the extraction solvent.

results. On the other hand, the ability of varying spectral patterns by changing solvent conditions can be advantageous. Since bacterial identification using the MALDI method is based on one or a set of characteristic peaks in the spectrum, once the biomarkers are identified for a particular bacterium, optimal sample/matrix preparation methods can be designed for the sensitive detection of these biomarkers. While a universal sample preparation method for detecting all biomarkers for many different bacterial strains and/or species is desirable, such a preparation protocol may be difficult to find. The utility of multiple sample preparation protocols with each optimized for a small number or a group of biomarkers may be a sensible approach. Multiple sample handling is a standard feature in commercial MALDI instruments. Automation of the sample preparation step is possible. With this in mind, we can again examine the MALDI spectra of *E. coli* 9637 shown in Figure 2.1. The peaks shown around m/z 7709 and 3854 are the dominant peaks in Figure 2.1B. If these peaks were to be used for the identification of *E. coli*, the experimental conditions used for generating Figure 2.1B are preferred to those used for Figure 2.1A or 2.1C, from the MALDI detection point of view.

2.3.2 Effect of the Salt Content

Another important factor influencing the detection of peptides and proteins is the salt content of the bacterial samples. Figures 2.2A and B show the mass spectra of *B. thuringiensis* 19267 obtained with and without the use of membrane filtration, respectively. (For membrane desalting, Millipore cellulose ester filters having 0.025 μm pore size were used.) A greater number of peaks are detected in Figure 2.2A, demonstrating that salt content in the sample can have a significant effect on signal detection. The presence of a large amount of salt in a sample can affect the efficiency of analyte incorporation in crystals during sample preparation, the analyte desorption properties, and/or the ionization efficiency for certain peptides and proteins.

In Figure 2.2, the peaks with similar masses found in both spectra are labelled with m/z values. Several pairs of peaks at m/z's 2788/2793, 3683/3688, 5580/5588, 7185/7189, 7231/7236, 7372/7375, and 9521/9523 are likely from the same peptides and proteins. The mass discrepancy observed in the corresponding peaks in the two spectra can be attributed to the difficulty of defining the peak centroid due to the low resolving capability of the instrument used, as well as the mass calibration (external calibration was used). Note that all the peaks detected in Figure 2.2B are observed in the spectrum shown in Figure 2.2A; but the relative intensities of these peaks are different. This suggests that many more peptides and proteins are present in the extract than those detected in these MALDI spectra. The variation of salt content from sample to sample can result in the detection of some peptides and proteins preferentially over the others, resulting in different mass spectral patterns. This underscores the importance of searching for unique biomarkers via their observed masses, instead of relying on the spectral pattern differences for bacterial identification and differentiation.

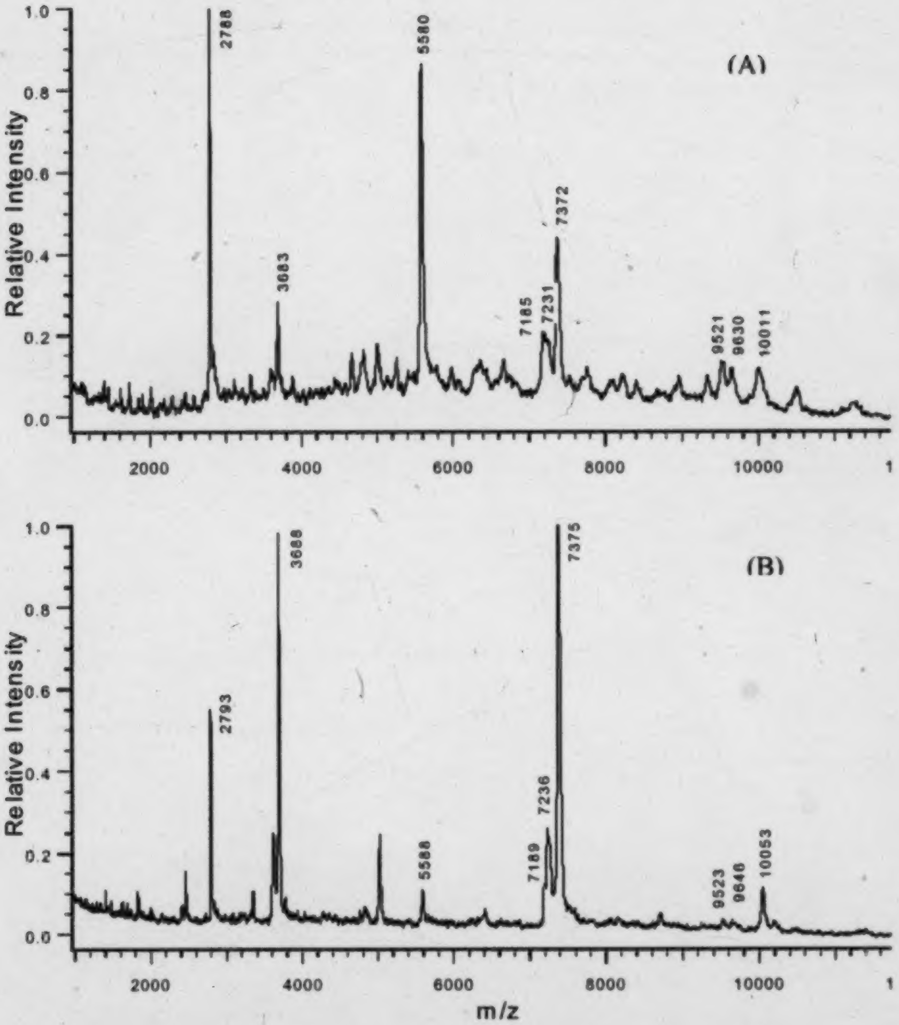

Figure 2.2 MALDI spectra of *B. thuringiensis* 19267 obtained at the ERDEC lab. (A) The bacterial extract was desalted by a membrane filter prior to the MALDI sample preparation, (B) no desalting step was used. A 50-mM ammonium bicarbonate solution was used for protein extraction.

2.3.3 Effect of the Extraction Solvent

In addition to the aforementioned issues related to sample preparation in MALDI, the method of protein extraction from bacterial samples can also have a major impact on mass spectral patterns. Using the solvent suspension method for protein extraction, it was found that mass spectral patterns can be significantly affected by the type of suspension solvent used. This may be due to the solubility differences of the peptides and proteins in different solvents and/or due to the differing extents to which the suspension solution may lyse or perforate the cells. Figure 2.3 shows the mass spectra of *E. coli* obtained with two different extraction solvents, but under the same matrix/sample preparation condition. Figure 2.3A was obtained by using 0.1% TFA as the suspension solvent for protein extraction. The second-layer solution consists of the FMW solvent mixture and the sample solution in 0.1% TFA (1:1) (all by volume). For Figure 2.3B, the FMW solvent mixture was used for extraction. The second-layer solution for Figure 2.3B was prepared by adding an appropriate amount of 0.1% TFA to the FMW extract so that the final composition of the solution is the same as that used for Figure 2.3A. Thus, any mass spectral pattern differences observed under these conditions are solely due to the effect of the type of extraction solvent.

Figure 2.3 clearly shows that the extraction solvent can have a significant effect on the spectral patterns. There are several common peaks present in the spectra shown in Figures 2.3A and 2.3B. They are likely from the same peptides and proteins dissolved in both solvents. Comparing these two spectra also reveals that different proteins are extracted using the two solvent systems. This example illustrates that the selection of the type of extraction solvent is important for detecting the potential biomarkers. In addition to TFA and the FMW solvent mixtures, several other extraction solvents were tested.

Figure 2.4 shows the mass spectra of *E. coli* 9637 obtained using various solvents for extraction. Pure methanol (Figure 2.4 A) and isopropanol (Figure 2.4 B) were found to be poor solvents for extraction. Water, ammonium bicarbonate, and Tris-HCl buffer were found to provide similar spectra for *E. coli* 9637, as shown in Figures 2.4 C-E.

In using any aqueous solution for extraction, it was found that the change in solution pH has a profound impact on mass spectral pattern. This is illustrated in Figure 2.5 for three spectra of *B. thuringiensis* 19267 obtained by using ammonium bicarbonate solutions with different pH values for extraction. In this experiment, the pH of the 50 mM ammonium bicarbonate solution was adjusted by adding a few drops of 1 M HCl. Figure 2.5 shows that only a small change in pH can result in very different mass spectra. The extent of the pH effect was found to be sample dependent. For example, Figure 2.6 shows the mass spectra of *B. thuringiensis* 10792, a different strain from *B. thuringiensis* 19267. In this case, the overall pH effect is not as dramatic as those shown in Figure 2.5. However, a significant change in mass spectral pattern is noted when the extraction solution pH is changed from pH 8.5 (Figure 2.6C) to pH 8.0 (Figure 2.6B). The observed pH effect is likely a reflection of the variation of peptide/protein solubilities in different solutions during extraction.

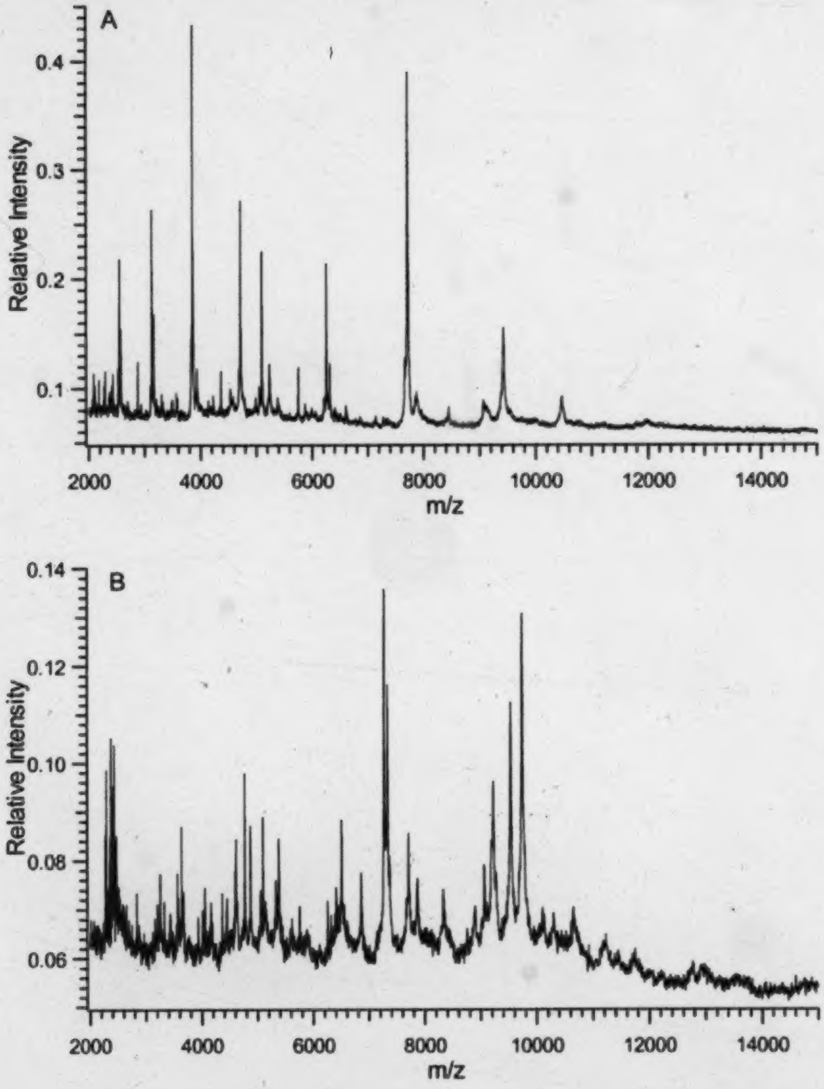

Figure 2.3 MALDI spectra of *E. coli* 9637 obtained at the U of A lab under a controlled experimental condition to illustrate the effect of the type of extraction solvent on mass spectral pattern. (A) 0.1% TFA was used as the extraction solvent and (B) a mixture containing 17% formic acid, 33% methanol, and 50% water (by volume) was used as the extraction solvent (see text for details on performing this extraction). The final solvent composition used for preparing the second-layer solution in MALDI was adjusted to be the same in both cases.

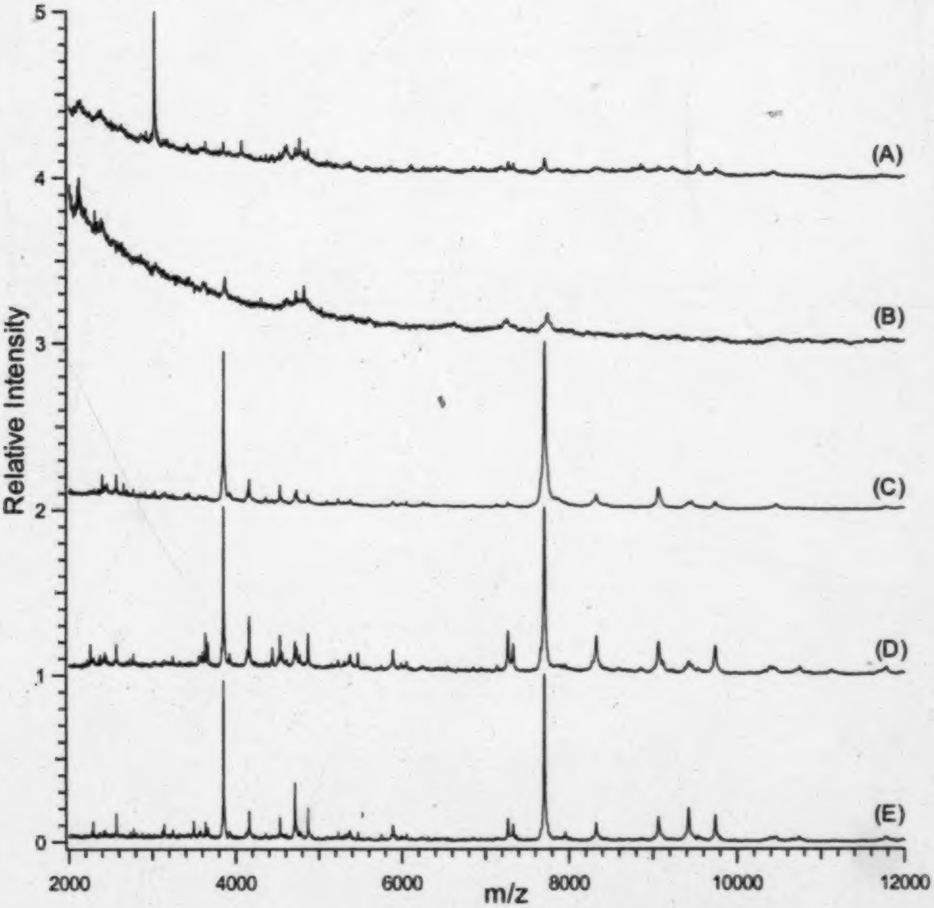

Figure 2.4 MALDI spectra of *E. coli* 9637 obtained at the U of A lab by using (A) methanol, (B) isopropanol, (C) water, (D) ammonium bicarbonate, and (E) Tris-HCl buffer as the extraction solvent. The second-layer solution in the two-layer sample preparation consisted of 17% formic acid/33% isopropanol/50% water and the bacterial extract.

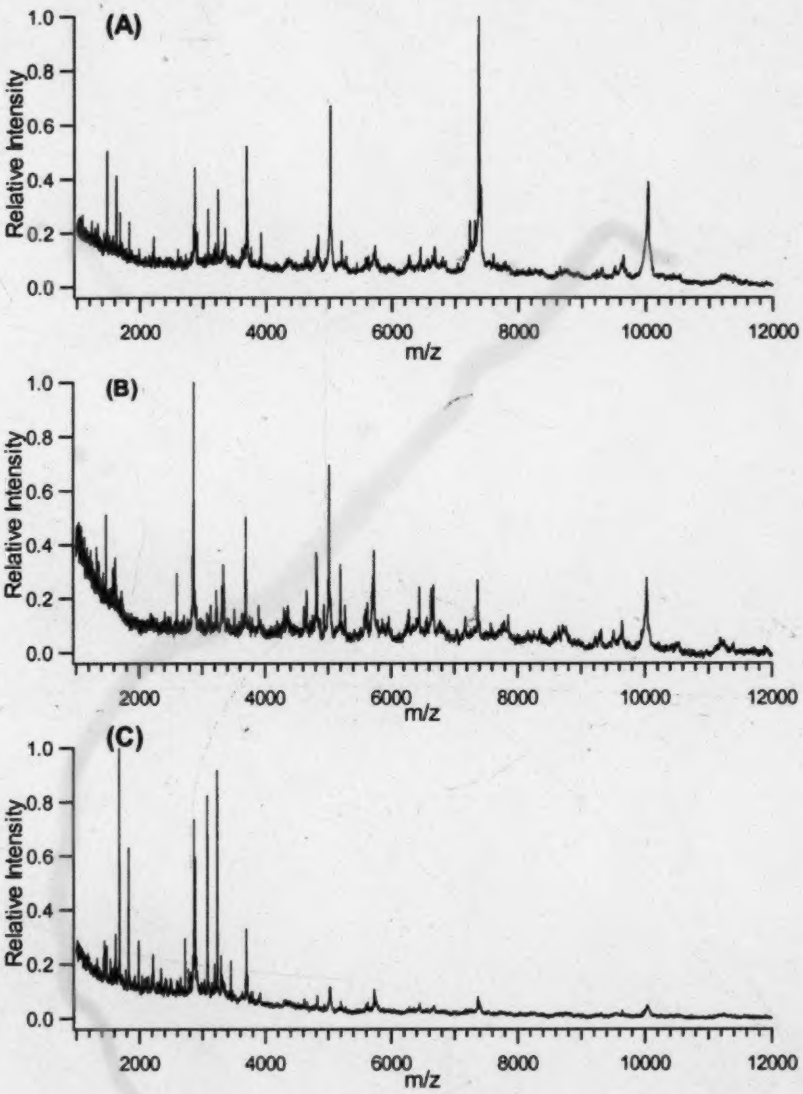

Figure 2.5 MALDI spectra of *B. thuringiensis* 19267 obtained at the U of A lab by using ammonium bicarbonate solutions with different pH values: (A) pH=7.6, (B) pH=8.0, and (C) pH=8.5.

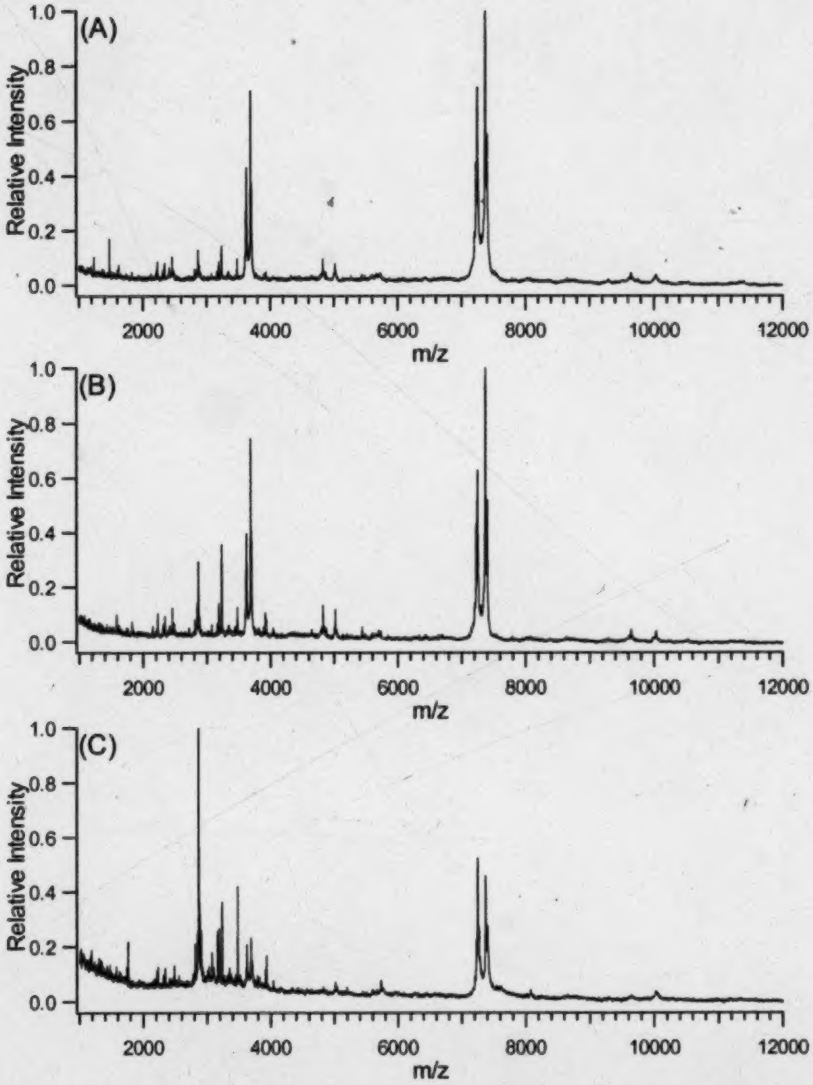

Figure 2.6 MALDI spectra of *B. thuringiensis* 10792 obtained at the U of A lab by using ammonium bicarbonate extraction solutions with different pHvalues: (A) pH=7.6, (B) pH=8.0, and (C) pH=8.5.

It is clear that the type of extraction solvent and, in the case of aqueous solution, the pH of the solution can have a major impact on the mass spectral pattern. These two parameters should be well controlled to obtain reproducible results. From the method development point of view, one can potentially vary these two parameters to achieve optimal conditions for sensitive and selective detection of biomarkers.

2.3.4 Effect of the Extraction Method

In the reported studies,[1-4] different protein extraction methods were used. We have compared the results obtained from one of the reported extraction methods[2] with the solvent suspension method described in this work. Figure 2.7 shows the mass spectra of *E. coli* 11775, a different strain from *E. coli* 9637, obtained by the two extraction methods. Figure 2.7A was obtained at the U of A lab using the same experimental condition as that used in Figure 2.1B for *E. coli* 9637. Figure 2.7B shows the mass spectrum obtained at the ERDEC lab using a modification of the chemical lysis based extraction procedure of Ref 2. Dry, lyophilized bacteria (2-3 mg) were placed in a microcentrifuge tube to which was added 75 µl each of the following solutions: 10 mM Tris buffer (pH 8.0), 1% SDS, and 0.1 mM β-mercaptoethanol. This mixture was vortexed, then incubated at 95 C for 20 minutes. The cloudy solutions were cooled to room temperature, 10 µL DNAse(I) (1 mg/mL) was added, and the solutions were held at room temperature for 20 minutes. The mixture was centrifuged at low speed for 10 minutes, and the clear supernatant was transferred to a clean microcentrifuge tube. The soluble proteins were precipitated by addition of 1 mL of cold methanol followed by storage at 0 C for 20 minutes. The proteins were pelleted by centrifugation at high speed for 10 minutes. The supernatant was removed and the precipitate was air dried for 10 minutes. Eight samples were reconstituted with 75 µL Tris (1 mM, pH 8.0), combined and dried using a Centrivap. For MALDI analysis, the dried precipitate was reconstituted in 100 µL water. The resulting solution was centrifuged at low speed to remove any

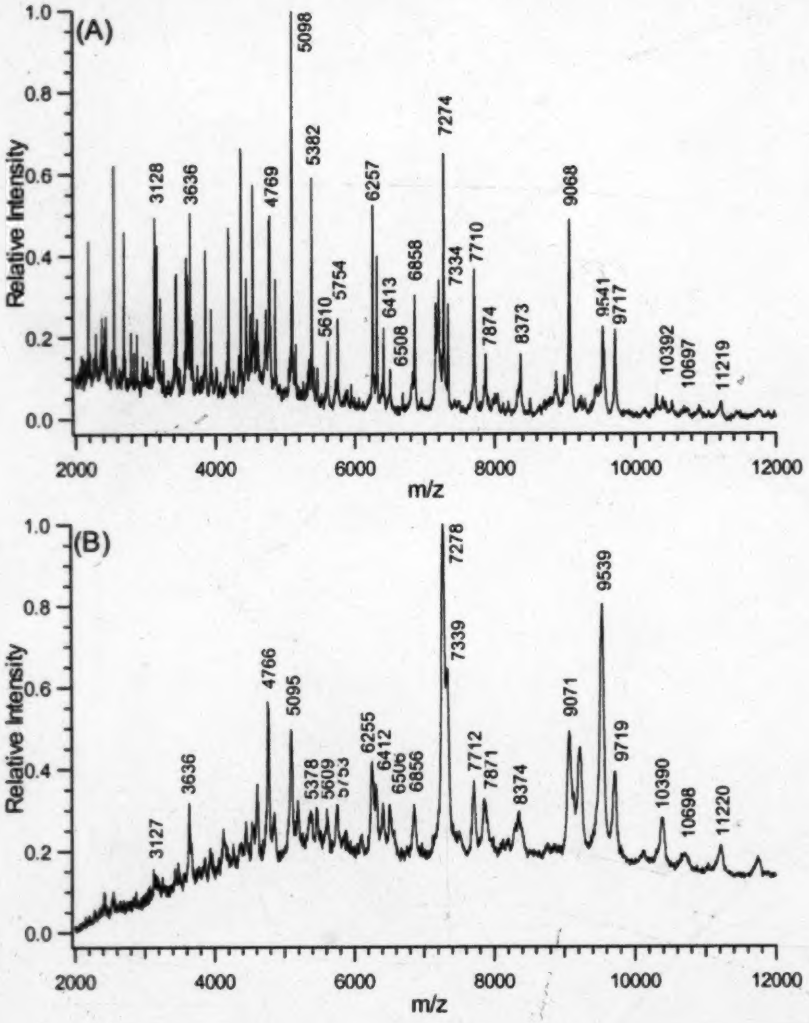

Figure 2.7 MALDI spectra of *E. coli* 11775 obtained by using different extraction methods: (A) using the solvent suspension method at the same condition as that of Figure 2.1B (collected at the U of A lab) and (B) using the enzyme-based extraction method (collected at the ERDEC lab).

particulates before use. The MALDI samples were prepared by adding 1 μL of the extract solution to 9 μL of matrix solution (saturated HCCA in 1:1 acetonitrile/0.3% TFA by volume). This solution was vortexed for 30 s and 1 μL was immediately spotted onto a clean MALDI sample pin and allowed to air dry.

As Figure 2.7 shows, most peaks detected in Figure 2.7B are also found in Figure 2.7A, despite the fact that completely different sample preparation procedures were followed and different MALDI instruments were used to obtain the two spectra. Peaks observed in common between the two approaches are labelled by the m/z values and they comprise a very large fraction of the observed peaks, particularly above m/z 4500.

It is interesting to make a comparison of spectral patterns obtained for the two *E. coli* strains. The mass spectral patterns shown in Figure 2.7 for *E. coli* 11175 are quite different (for example, many more peaks observed) from those of *E. coli* 9637 shown in Figure 2.1 and in Figure 2.3A, where nominally the same sets of peaks are observed though intensities vary due to sample preparation conditions. This comparison would suggest that the mass spectral patterns reveal a difference between strains. However, when Figure 2.3B is considered, one finds a considerable similarity between the prominent mass spectral peaks for *E. coli* 11775 (Figure 2.7) and for *E. coli* 9637 extracted by the FMW formulation (Figure 2.3B). These comparisons suggest that, while the application of a given extraction approach may appear to reveal differences due to strains, the application of alternative extraction approaches may cause additional peaks to appear. When one considers all the peaks appearing from different extraction approaches, the (initially) apparent strain-based spectral differences may become difficult to discern and certainly difficult to verify as strain-based differences, given the dependence of spectra on extraction approach, sample preparation, and content of salt and other common biological solution constituents.

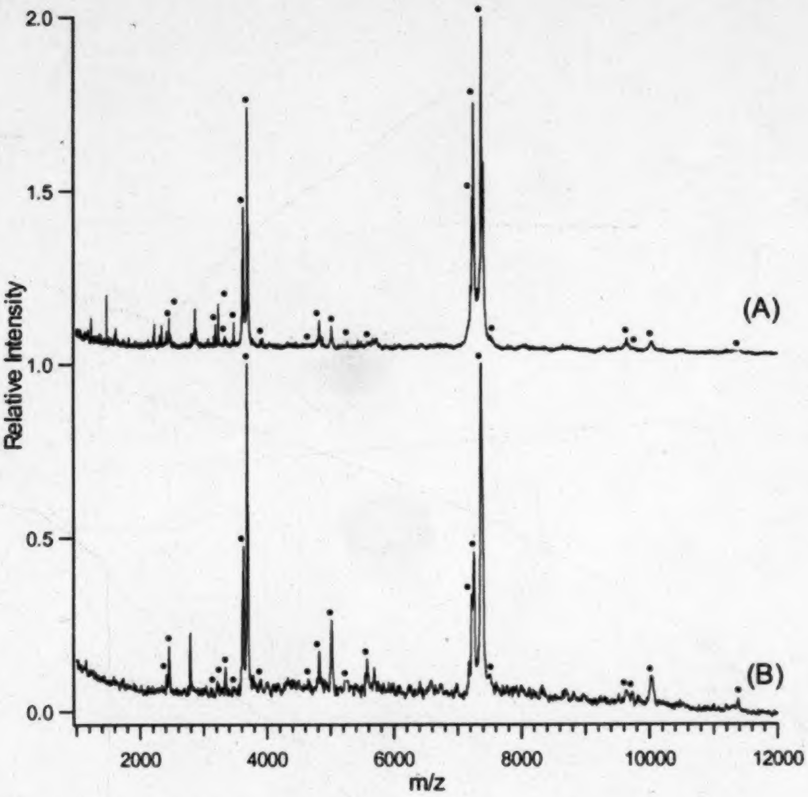

Figure 2.8. MALDI mass spectra of *B. thuringiensis* 10792 obtained at U of A (Figure 4.8A) and ERDEC (Figure 4.8B). Common peaks are designated by solid circles. Ammonium bicarbonate solution was used for extraction and the two-layer method was used for sample/matrix preparation.

2.3.5 Inter-laboratory Comparison

Although the experimental conditions can have significant effects on mass spectral patterns, it was found that reproducible results can be readily obtained within a laboratory with appropriate controls on the conditions applied. We also examined the interlaboratory reproducibility. Figure 2.8 shows the mass spectrum of *B. thuringiensis* 10792 obtained at the University of Alberta and ERDEC using the ammonium bicarbonate suspension. The common peaks detected in the two panels of Figure 2.8 are marked with solid circles. It can be seen that most of the observed peaks are replicated in the two studies for *B. thuringiensis* 10792. In this case, even the relative intensities of many of these peaks are similar in both spectra. However, we also found that, for some bacteria, the relative intensities can be quite different for the spectra obtained in two labs using the same protocol for sample extraction and MALDI preparation. An example is shown in Figures 2.9A and 2.9B for *B. thuringiensis* 19267 using the ammonium bicarbonate suspension. For comparison, Figure 2.9C shows the spectrum of *B. thuringiensis* 19267 obtained by using an extraction method involving sequential addition of formic acid, methanol, and water (see Figure caption). As Figures 2.9A and 2.9B show, the spectral patterns at the m/z region between 1000 to 4000 are quite different; but by far most of the peaks observed are common between the two spectra, as marked by solid circles. In light of the above discussion considering the factors affecting the spectral patterns, it is not totally surprising that varying intensities are observed among the peaks recorded in different laboratories. The solvents, matrices, glassware, and the sample probe can also introduce salts or impurities that may affect the spectral patterns. The intensity of the laser radiation at the sample is also difficult to duplicate in different labs.

The ability of achieving reproducible results that can be replicated in different laboratories is certainly important to establish the validity of the MALDI technique for bacterial identification. This work illustrates that control of the experimental parameters that can be readily controlled does permit a substantial, though not complete, replication of bacterial peptide/protein extract spectra. Common masses observed by the two labs are provided in Table 2.1 for *B. thuringiensis* 10792 and in Table 2.2 for *B. thuringiensis* 19267.

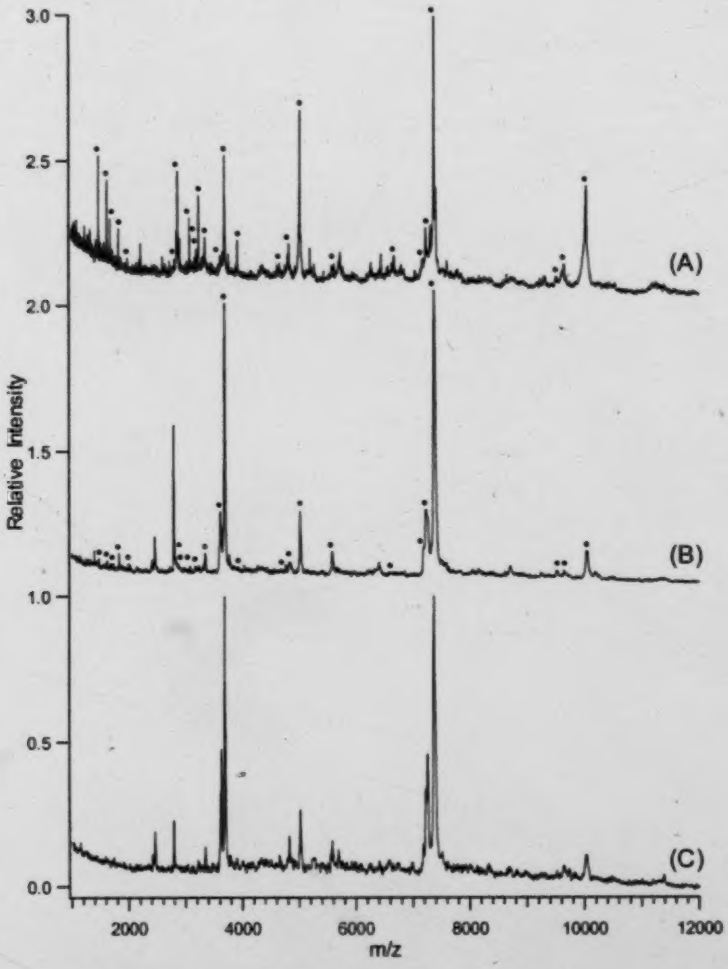

Figure 2.9 MALDI mass spectra of *B. thuringiensis* 19267 obtained at U of A (Figure 2.9A) and at ERDEC (Figure 2.9B). Common peaks are marked with solid circles. Ammonium bicarbonate solution was used for extraction and the two-layer method was used for sample/matrix preparation. In panel 2.9C is shown the mass spectrum obtained at ERDEC using the formic acid (F)/methanol (M)/water (W) extraction, wherein the major peaks are essentially replicates of those in panels 2.9A and 2.9B. To clarify, the approach used for Figure 2.9C involved sequential addition of the solution components in order F, M, W with vortexing between additions of components. If the FMW solution is prepared first, then added to the sample with vortexing, a significantly different intensity distribution is observed.

2.3.6 Comparison with Related Efforts

The notion of species, genus, and strain specific protein biomarkers has been invoked to utilize protein MALDI mass spectral profiles for bacterial differentiation/identification. This notion is certainly well-founded in principle due to the fact that the genomic structures of organisms determine the differences between these organisms and their classification according to species, genus, and strain.[16] Since the genomic sequences code for the proteins in the organism, proteins reflect the genomic differences between organisms and mass spectral examinations of proteins should in principle serve as a basis for differentiation. Specifically, among members of a genus, such as *Bacillus*, some common masses (genus-specific biomarkers) would be expected. Similarly, different strains of a species, such as *B. thuringiensis*, should have common masses, not observed for other *Bacillus*, that are species-specific. In addition, different species of the same genus should have some different masses among them. A significant degree of similarity should therefore be expected, certainly among members of a species, if a "statistically significant" number of proteins present in a bacterium are observed.

With this background, it should be possible to make comparisons of the results presented in this work with and among the results reported by other workers. Several previous studies addressing MALDI of *Bacillus* species proteins have been conducted, using either direct analysis of extracted proteins[2] or of whole cells mixed with matrix[4,17] as well as chromatographic separation of extracted proteins followed by off-line MALDI analysis of the chromatographic fractions.[18] We believe that the application of variant experimental conditions, in extraction and/or sample handling and preparation, accounts for the considerably higher level of disagreement than of agreement among these studies.

Specifically, no common prominent ions were observed (by different laboratories) to which attribution as species- or genus-specific biomarkers can be made with confidence. Similarly, a comparison among the results of this study and others on various strains of *E. coli*[3,18,19] presents difficulties with regard to finding commonality among prominent mass peaks. These comparisons cannot, by any means, suggest that any observations to date are "correct" or "incorrect". The capability to attain good reproducibility within each laboratory is evident in each study. Rather, this comparison suggests that application of variant experimental approaches results in the observation of different subsets of the total set of proteins available in the sample. It further demonstrates that the mass spectrometric analysis of bacterial proteins is indeed a complex analytical challenge and optimization of sample treatment/analysis protocols is clearly required. Lubman and coworkers[18] demonstrated that application of a separation step such as liquid chromatography prior to off-line MALDI analysis generates an overall larger number of peptide/protein masses observed. This can be attributed to fewer analytes in each fraction, hence, less ion suppression in MALDI. Such an approach should enhance the probability of identifying common masses where expected. Unfortunately, the mass resolution applicable to that effort[18] does not permit truly valid mass matches to be identified.

The significance of the present work is the demonstration of a capability to attain substantial replication of results by two different laboratories when common sample treatment approaches are used, in addition to the identification of some factors that have the potential to cause variability in observed spectra. When optimization of sample treatment/analysis methodologies is accomplished, it will become possible to begin to identify with confidence species-, genus-, and strain-specific protein biomarkers through the development of sufficiently large libraries of bacterial protein mass spectral data.

2.4 Conclusions

This work demonstrated that spectral reproducibility in MALDI analysis of peptides and proteins directly from bacterial extracts can be influenced by a number of experimental factors. With regard to sample preparation, it has been demonstrated that the solvent composition in preparing the MALDI matrix/sample solution and the salt content in the sample can have a significant effect on mass spectral pattern. On the issue of extraction approach, we find that the solvent suspension method provides a rapid means of extracting peptides and proteins from bacterial samples. However, different mass spectra may be obtained using different protein extraction processes. The type of extraction solvent and the pH of an aqueous solution used for extraction can have a major impact on observed spectra.

Using the same optimized sample extraction/preparation strategies, we found that substantial reproducible mass spectra can be obtained, suggesting that the technique has the potential to be a valuable bacterial differentiation/identification tool, once optimum sample extraction/preparation strategies have been developed that translate well from one laboratory to another. It can be concluded that despite the significant variation of mass spectral patterns that may result from minor changes in experimental conditions, there are many peaks whose detection seems assured even with some variation in specific processes used. These "conserved" peaks represent the ones that have the highest potential for use as biomarkers for bacterial identification.

Table 2.1 The m/z values of the common peaks observed for *B. thuringiensis* 10792 from two labs.

Figure 2.8A	Figure 2.8B
2418	2419
2460	2457
3216	3216
3234	3237
3345	3344
3479	3485
3626	3629
3688	3685
3898	3897
4646	4647
4824	4822
5017	5016
5267	5269
5598	5598
7230	7223
7249	7256
7373	7368
7517	7512
9648	9652
9743	9746
10038	10036
11388	11385

Table 2.2 The m/z values of the common peaks observed for *B. thuringiensis* 19267 from two labs.

Figure 2.9A	Figure 2.9B
1477	1477
1624	1627
1681	1684
1828	1832
1986	1990
2835	2831
2867	2873
3077	3083
3166	3171
3192	3196
3235	3241
3347	3347
3616	3616
3688	3688
3920	3928
4659	4759
4825	4826
5018	5020
5582	5588
6678	6682
7183	7189
7232	7236
7374	7375
9521	9523
9650	9646
10036	10053

2.5 Literatures Cited

1. Cain, T. C.; Lubman, D. M.; Weber, Jr., W. J. *Rapid Commun. Mass Spectrom.* **1994**, 8, 1026.
2. Krishnamurthy, T.; Ross, P. L.; Rajamani, U. *Rapid Commun. Mass Spectrom.* **1996**, 10, 883.
3. Holland, R. D.; Wilkes, J. G.; Rafii, F.; Sutherland, J. B.; Persons, C. C.; Voorhees, K. J.; Lay, Jr., J. O. *Rapid Commun. Mass Spectrom.* **1996**, 10, 1227.
4. Krishnamurthy, T.; Ross, P. L. *Rapid Commun. Mass Spectrom.* **1996**, 10, 1992.
5. Dai, Y.; Whittal, R. M. ; Li, L. *Anal. Chem.* **1996**, 68, 2494.
6. Whittal R. M.; Li, L. *Anal. Chem.* **1995**, 67, 1950.
7. Whittal, R. M.; Russon, L. M.; Weinberger, S. R.; Li, L. *Anal. Chem.* **1997**, 69, 2147.
8. Whittal R. M.; Li, L. *American Laboratory*, **1997**, 29, 30.
9. Strupat, K.; Karas, M.; Hillenkamp, F. *Int. J. Mass Spectrom. Ion Processes*, **1991**, 111, 89.
10. Weinberger, S. R.; Boernsen, K. O.; Finchy, J. W.; Robertson, V.; Musselman, B. D. *Proceedings of the 41st ASMS Conference on Mass Spectrometry and Allied Topics*; San Francisco, CA, 1993; pp 775a-b.
11. Xiang F.; Beavis, R. C. *Rapid Commun. Mass Spectrom.* **1994**, 8, 199.
12. Vorm, O.; Roepstorff, P.; Mann, M. *Anal. Chem.***1994**, 66, 3281.
13. Billeci T. M.; Stults, J. T. *Anal. Chem.* **1993**, 65, 1709.
14. Gusev, A. I.; Wilkinson, W. R.; Proctor, A.; Hercules, D. M. *Anal. Chem.* **1995**, 67, 1034.

15. Cohen S. L.; Chait, B. T. *Anal. Chem.* **1996**, 68, 31.

16. Alberts, B.; Bray, D.; Lewis, J.; Raff, M; Roberts, K.; Watson, J. D. *Molecular Biology of the Cell*, 3rd Ed., Garland Publishing Inc., New York, 1994.

17. Wahl, K. L.; Valentine, N. B.; Gantt, S. L.; Saenz, A.; Clauss, S. A.; Kingsley, M. T. In *Proceedings of Joint Services Workshop on Biological Mass Spectrometry*, ERDEC Special Publication, in press.

18. Liang, X. L.; Zheng, K. F.; Qian, M. G.; Lubman, D. M. *Rapid Commun. Mass Spectrom.*, **1996**, 10, 1219.

19. Arnold, R.; Houston, C.; Reilly, J. P. In *Proceedings of the 45th ASMS Conference on Mass Spectrometry and Allied Topics*, Palm Springs, CA, **1997**; pp31.

Chapter 3

Mass Spectrometric Methods for Generation of Low-Mass Proteome Database to be Used for Bacterial identification[a]

3.1 Introduction

Bacterial identification can be achieved by first generating a mass spectral profile of proteins from bacterial whole cells or cell extracts using MS, and then comparing the spectrum to the archived spectra of different individual bacteria.[1-4] This approach, however, greatly depends on the mass spectral reproducibility. Although the experimental conditions (for example, bacterial extraction and MALDI sample preparation conditions) can be well controlled to obtain reproducible spectra as discussed in Chapter 2,[5] it is very difficult to control the biology related factors such as cell growth and killing conditions.[6] An alternative approach involves detecting a subset of protein masses from an unknown bacterium and searching the set of masses against the public proteome database.[7] Although this approach does not rely on mass spectral reproducibility, it is very much dependent on the extent and quality of the database as well as the mass data obtained from the unknown.

In this chapter, we will demonstrate that current public proteome database has several limitations which constrain its applicability to bacterial identification. We are confronted with the necessity of generating protein mass databases tailored for bacterial identification based on MS techniques. This can be done by a combination of HPLC separation with MALDI and ESI MS. In this work, issues related to the generation of protein mass databases based on protein mass analysis by MS will be addressed. The applicability of such protein mass databases will be evaluated.

---

[a] A form of this chapter is ready to submit as: Z. Wang, K. Y. Dunlop, L. Li *"Mass Spectrometric Methods for Generation of Low-Mass Proteome Database to be Used for Bacterial identification"*, Mr. Kevin Y. Dunlop collected the ESI data.

3.2 Experimental

3.2.1 Bacterial Protein Extraction

Bacteria samples were from ERDEC as described in Section 2.2.1. Bacterial extracts were prepared by solvent suspension method as described in Section 2.2.2. About 25 mg of lyophilized bacterial cells (*Escherichia coli* 9637, *Bacillus megaterium* and *Citrobacter freundii*) were suspended in 1 mL 0.1% TFA, vortexed for about 3 min and centrifuged at 14000 rpm for 5 min. The supernatant was then removed into a fresh vial. This extraction process was repeated 3-5 times to maximum the extraction efficiency. The supernatants were pooled and filtered using a Mirocon-3 with 3000 Da molecular mass cut-off (Millipore, Bedford, MA), and then concentrated to about 0.5 mL by Speed-Vac.

3.2.2 HPLC Fractionation

Solvent delivery and separations were performed on a Hewlett-Packard (Palo Alto, CA) HP1100 HPLC system. Separations were optimized for each bacterial extract and the conditions are listed in Table 3.1. For online LC/ESI, 40 μL of bacterial extract was separated on a 150 x 2.1 mm i.d. C_8 column (5 μm particles with 300Å pore size, Vydac, Hesperia, CA) at a flow rate of 200 μL/min. For LC/off-line MALDI, 100 μL of bacterial extract was separated on a 250 x 4.6 mm i.d. Vydac C_8 column at a flow rate of 500 μL/min. The fractions were collected every minute using a Gilson FC 203B fractionation collector (Gilson, Middleton, WI).

Table 3.1 Optimized gradient separation conditions (shown as % solvent B. Solvent A: 0.05 %TFA in water; solvent B: 0.05% TFA in acetonitrile).

E. coli	*B. megaterium*	*C. freundii*
0 min – 2%	0 min – 2%	0 min – 2%
10 min – 20%	10 min -- 20%	10 min – 20%
40 min – 40%	40 min – 40%	30 min – 40%
45 min – 55%	45 min – 55%	60 min – 55%
60 min – 90%	60 min – 90%	

3.2.3 Mass Spectrometry

3.2.3.1 ESI

The HPLC effluent was analyzed with a HP 1100 MSD quadrupole mass spectrometer. The m/z range from 500 to 3000 was scanned in 1.90 s and the ions were detected with a high-energy dynode detector. Control of both the HPLC and MS systems was accomplished with HP ChemStation software. To compensate for the signal suppressing effect from Trifluoroacetic acid (TFA) in the mobile phase during ESI analysis, glacial acetic acid (HPLC grade, Fisher Scientific, Fair Lawn, NJ) was added to the column effluent at 100 μL/min by a PEEK "Y" connector and a syringe pump (Cole Parmer, Vernon Hills, IL). The "Y" was connected to the electrospray interface by a 30 cm piece of PEEK tubing (0.005 inch i.d.). It is found that about 10 times signal enhancement can be achieved by the post-column addition of acetic acid. Early results showed the total ion chromatogram (TIC) intensities varied according to the voltage at the capillary exit. A large voltage drop in the region between the capillary exit and the first skimmer produced more intense signals in the TIC. However, it also had the effect of stripping charges from the analytes as well as producing severe fragmentation due to collision-induced dissociation (CID). In order to overcome the detrimental CID effects caused by a large and constant fragmentation voltage, the voltage was varied as the quadrupole scanned across the selected mass range. The optimized voltage ramp was found to be: m/z 500 = 60V, 1000 = 120V, 3000 = 220V. The resulting mass spectra showed a greater number of peaks and a higher signal to noise ratio compared to those obtained using a constant fragmentation voltage.

3.2.3.2 MALDI Analysis

The MALDI results were obtained in a time-lag focusing MALDI-TOF mass spectrometer which has been described in detail elsewhere.[8] A 337-nm and 3-ns pulse width of laser beam from a nitrogen laser (model VSL 337ND, Laser Sciences Inc., Newton, MA) was used for desorption. In general 50-100 laser shots (3-5 μJ pulse energy) were averaged to produce a mass spectrum. Spectra were acquired and processed with Hewlett-Packard supporting software and reprocessed with the Igor Pro software package (WaveMetrics, Inc., Lake Oswego, OR).

For direct MALDI analysis, about 1 mg of lyopholized bacterial sample was extracted by 500 μL 0.1% TFA. For LC/off-line MALDI analysis, the HPLC fractions were concentrated by 50 times to about 10 μL before mixing with matrix for analysis. A two-layer method was used for MALDI sample preparation. α-cyano-4-hydroxycinnamic acid (HCCA) was used as the matrix. About 1 μL of 0.1 M HCCA in acetone/water (99/1, by volume) was applied to the MALDI probe tip to quickly form the first layer. For the second layer, the sample solution was mixed 1:1 with saturated HCCA in formic acid/isopropanol/water (1/2/3, by volume). About 0.6-1 μL of the second layer solution was then applied onto the first layer and allowed to dry. On-probe washing of the MALDI sample with water was performed to remove the salts. External mass calibration was done in the mass range of 2-20 kDa using insulin chain B, horse cytochrome c and its multiply charged species and dimer.

3.3 Results

The simplest database for bacterial identification will likely consist of protein mass tables reflecting the proteome of the individual bacteria. In the following sections, mass tables obtained using MS techniques for the detection of low-mass (2–20 kDa) bacterial proteins are presented. *E. coli* was chosen as a model bacterium for further investigation since its proteome has been widely studied, which makes it possible to tentatively identify proteins based on protein masses alone, especially for low mass proteins. Identification of possible protein biomarkers will form a scientific base on the validity of the protein mass tables created by MS for bacterial identification. For example, the peaks observed by MS are from bacterial proteins, not from possible contaminants that may be introduced during cell growth or sample preparation.

3.3.1 LC/ESI-MS

Figure 3.1 shows a typical TIC of a cell extract obtained by LC/ESI-MS. The unusual chromatographic peaks were purposely produced by removing the static mixer in the HPLC pump. The oscillating nature of the ion current created from the removal of the mixer helps the automated mass spectral integration and interpretation.[9] The protein m/z values detected from the three bacterial samples are shown in Tables 3.2 to 3.4. The bolded m/z values correspond to those found using LC/off-line MALDI and the underlined are those whose molecular masses match with proteins in the Swiss-Prot or TrEMBL databases. The mass accuracy from this instrument is typically better than 0.02%.

Table 3.2 $(M+H)^+$ from *E. coli* extract by online LC/ESI-MS.

2008	**2736**	5171	**7274**	**9226**	**9740**	**15693**
2123	**3510**	**5550**	**7707**	**9231**	**10386**	15767
2139	**3625**	**6255**	**7782**	**9265**	**10459**	
2375	**3793**	**6316**	**7855**	**9518**	11224	
2416	4481	6330	**9064**	**9536**	11297	
2432	**5037**	**7140**	**9192**	**9610**	**11782**	
2505	**5097**	**7272**	**9209**	9684	**13094**	

Table 3.3 $(M+H)^+$ from *B. megaterium* extract by online LC/ESI-MS.

3048	**6336**	7157	**7711**	9754	**11063**
3187	6352	**7280**	**9332**	9768	**11538**
4379	6389	**7425**	9335	9829	11612
4741	**6393**	7451	**9349**	9884	**11726**
4816	**6449**	**7454**	**9352**	10026	12046
6262	**6578**	**7467**	**9620**	**10045**	12120
6276	**6897**	**7519**	**9694**	**10453**	12408
6316	**7111**	7649	**9747**	10696	

Table 3.4 $(M+H)^+$ from *C. freundii* extract by online LC/ESI-MS.

2063	3039	**4504**	**7752**	**10107**	**11870**
2391	**3237**	**5240**	**8278**	**10301**	11956
2431	**3450**	**5254**	8548	**10682**	**12157**
2473	**3565**	**5935**	9057	11162	12239
2515	**3581**	6530	**9196**	11252	16745
2676	**4007**	**6721**	9224	11675	**17178**
2793	**4421**	**7335**	**9523**	11691	
2853	**4437**	**7736**	**9527**	**11750**	

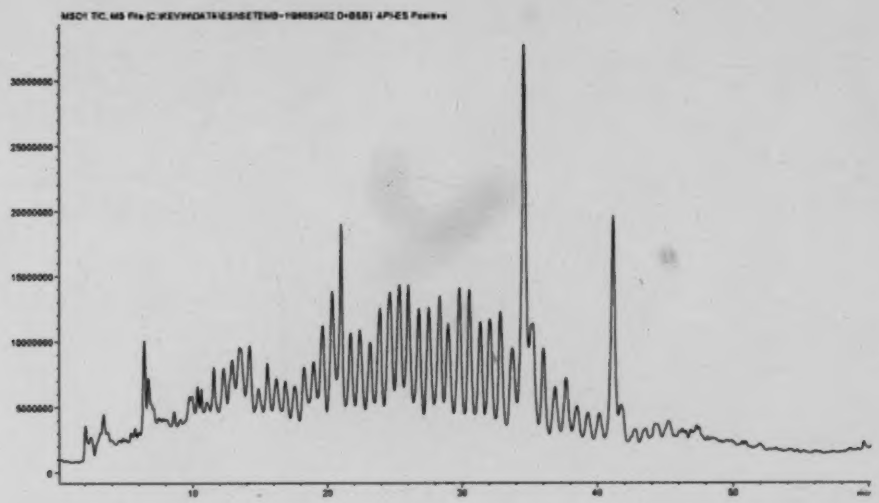

Figure 3.1 Total ion chromatogram of the *E. coli* extract separated by 2.1×150 mm C_8 column with the static mixer of HP1100 removed.

3.3.2 Direct MALDI

The MALDI mass spectra of the crude extracts from the three bacteria are shown in Figure 3.2. The m/z values of singly charged molecular ions are listed in Table 3.5-3.7. Insulin chain B and horse cytochrome c (multiply charged species and dimer) were used to externally calibrate the instrument over the mass range from 2000 to 20000 Da. The mass measurement accuracy is generally about 0.05%.

3.3.3 LC/Off-line MALDI

The protein m/z values detected by LC off-line MALDI are listed in Table 3.8-3.10. External mass calibration was done in the same manner as described above. The numbers of proteins detected by LC/off-line MALDI are 439, 286, and 157 for *E. coli*, *B. megaterium* and *C. freundii*, respectively. During MALDI analysis, clusters of peaks were often observed indicating the detection of methionine oxidized proteins in addition

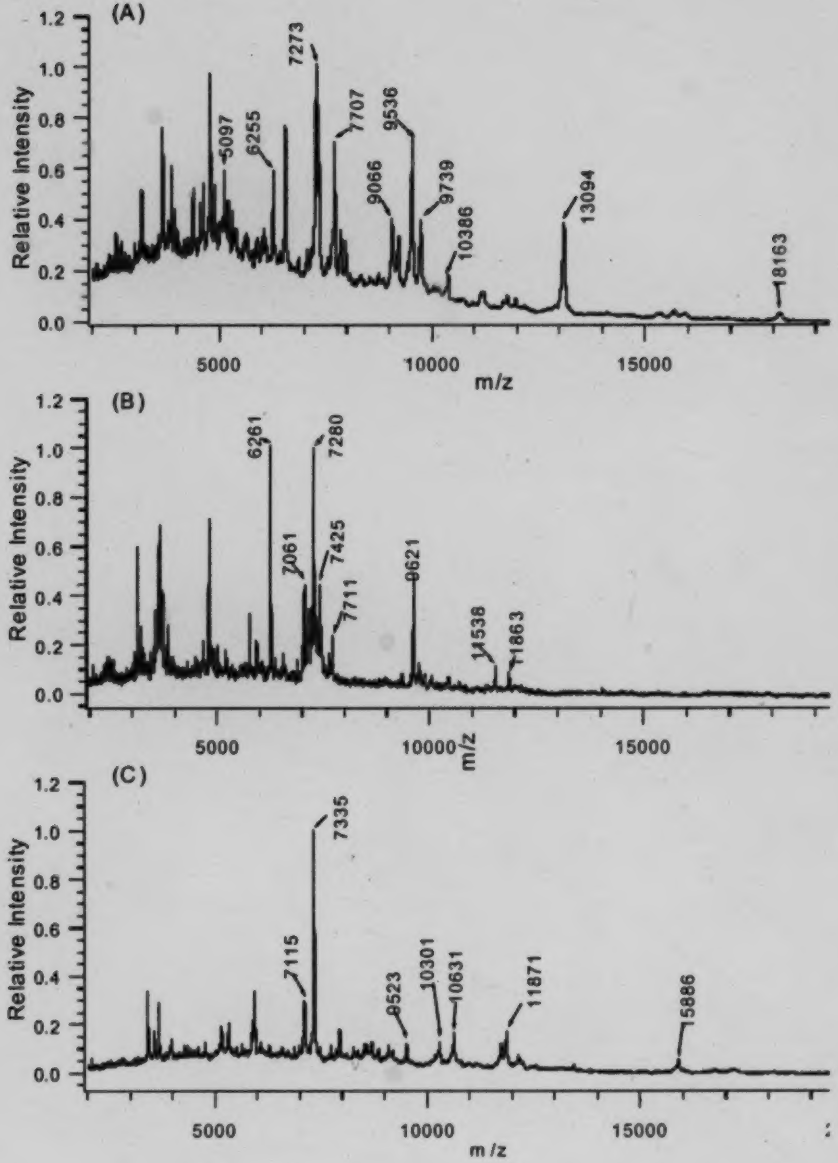

Figure 3.2 MALDI spectra of bacterial proteins extracted by 0.1% TFA aqueous solution. (A) *E. coli*, (B) *B. megaterium*, (C) *C. freundii*.

to their non-oxidized forms. The oxidation most probably occurred during sample storage and MALDI sample preparation. In the mass tables listed in this work, only the non-oxidized protein masses were included.

Table 3.5 $(M+H)^+$ from *E. coli* crude mixture by direct MALDI.*

2126	**5381**	**6413**	**7707**	**9536**	**11977**
2374	**5993**	**6853**	**7849**	**9739**	**13094**
2384	6057	**7061**	7970	**10386**	15682
5097	**6255**	**7273**	**9066**	**11206**	**18163**
5293	**6316**	**7333**	**9226**	**11783**	

Table 3.6 $(M+H)^+$ from *B. megaterium* crude extract by direct MALDI.*

6261	**7280**	**9351**	**10454**
6540	**7425**	**9621**	**11538**
6897	**7450**	9756	11863
7061	7648	9882	
7140	**7711**	**10044**	
7158	**9336**	**10418**	

Table 3.7 $(M+H)^+$ from *C. freundii* crude extract by direct MALDI.*

7090	**7736**	**8695**	10287	**11871**
7115	**7865**	8851	**10301**	12156
7141	**8278**	9107	**10631**	13448
7335	8547	**9523**	**11751**	**15887**

* Bolded masses match with the LC/off-line MALDI data, underlined masses match with the proteins in the Swiss-Prot or TrEMBL databases.

Table 3.8 $(M+H)^+$ from *E. coli* extract by LC/off-line MALDI.*

2004	2426	2846	**3510**	4445	5644	6669	7735	8997	10176
2006	**2431**	2858	3534	4447	5650	6684	7740	9048	10257
2009	2453	2860	3538	4456	5659	6700	7752	9055	10301
2014	2463	2874	3547	4465	5667	6703	**7781**	**9066**	10315
2026	2474	2878	3557	4473	5695	6721	7799	9080	10332
2041	2487	2884	3587	4484	5709	6725	**7852**	9160	10372
2044	2507	2910	3592	4500	5720	6772	7868	9175	10378
2050	2513	2970	3599	4562	5725	6787	8189	**9192**	**10387**
2054	2526	2975	3615	4716	5736	6827	8206	**9209**	10437
2067	2530	2984	**3624**	4722	5747	6854	8214	**9227**	10453
2075	2545	2999	3648	4792	5754	6866	8218	**9230**	**10463**
2085	2560	3003	3667	4895	5772	6890	8220	9252	10477
2106	2563	3014	3750	4903	5800	6942	8228	**9264**	10618
2123	2574	3022	3764	4940	5808	6951	8242	9281	10653
2127	2577	3037	3769	4985	5818	6958	8258	9296	10660
2132	2588	3054	3782	5006	5850	7060	8267	9369	10664
2141	2599	3073	**3793**	5017	5857	7070	8280	9424	10945
2166	2602	3090	3796	**5035**	5868	7095	8291	9431	11035
2197	2608	3101	3908	5040	5873	7109	8305	9439	11170
2208	2612	3119	3930	5052	5901	**7139**	8325	9458	11186
2232	2621	3127	3954	5073	5994	7158	8342	9475	11208
2237	2630	3135	3984	5087	6012	7169	8369	9478	11216
2240	2635	3159	3996	**5097**	6140	7185	8376	**9520**	11240
2260	2640	3192	4003	5144	6180	7255	8398	9527	11472
2264	2649	3212	4008	5154	6196	7265	8450	**9537**	11653
2279	2656	3226	4012	5180	6224	7269	8525	9545	11779
2282	2661	3251	4024	5202	6243	**7274**	8591	9555	**11783**
2293	2665	3259	4034	5253	**6255**	**7275**	8635	9573	11794
2297	2668	3286	4053	5295	6261	7281	8670	9583	11870
2301	2675	3289	4069	5362	6265	7283	8782	**9611**	11977
2305	2681	3295	4086	5378	6283	7290	8796	**9740**	12233
2312	2683	3302	4095	5382	6298	7293	8800	9750	12446
2331	2693	3309	4104	5396	**6316**	7298	8814	9754	12769
2339	**2737**	3324	4112	5414	6324	7307	8820	9766	13077
2353	2745	3331	4128	5430	6338	7314	8859	9785	**13095**
2360	2748	3389	4145	5449	6344	7321	8868	9834	13109
2374	2754	3401	4251	5467	6369	7326	8877	9852	13127
2376	2780	3406	4260	5470	6411	7334	8881	9885	13240
2382	2786	3417	4366	5480	6453	7412	8885	9952	13650
2386	2795	3424	4372	5493	6486	7570	8892	9982	14749
2395	2809	3476	4384	5549	6493	7600	8897	9996	14839
2400	2814	3486	4394	**5551**	6555	7616	8965	10045	**15693**
2403	2819	3492	4399	5566	6600	7619	8978	10106	18162
2417	2843	3500	4407	5585	6617	**7708**	8994	10123	

* Bolded masses match with the LC/ESI data, underlined masses match with the proteins in the Swiss-Prot or TrEMBL databases.

Table 3.9 (M+H)$^+$ from *B. megaterium* extract by LC/off-line MALDI.*

2023	2977	3543	4408	5016	5438	6362	7184	8720	10407
2166	2990	3555	4482	5027	5443	6382	7199	8794	10412
2211	2997	3629	4492	5030	5447	**6394**	**7280**	8818	10423
2223	3005	3637	4505	5040	5451	6414	7297	8849	10435
2232	3041	3694	4514	5055	5486	6446	7311	8862	**10453**
2248	3066	3701	4523	5059	5525	**6450**	7328	8866	10633
2270	3102	3752	4535	5070	5554	6492	7339	8942	10684
2281	3170	3796	4537	5073	5562	6498	7352	8980	**11063**
2290	**3187**	3800	4540	5090	5631	6508	7357	9016	11135
2296	3233	3811	4579	5095	5751	6563	**7425**	9189	11152
2313	3250	3817	4678	5106	5763	6566	7436	9328	11273
2325	3262	3822	4703	5113	5796	**6578**	**7453**	**9333**	11469
2354	3271	3839	4717	5139	5888	6580	**7467**	9348	11523
2364	3306	3879	4737	5143	5936	6591	7497	**9351**	**11539**
2421	3312	3935	**4742**	5147	5950	6595	7507	**9354**	11554
2435	3323	3941	4790	5165	5968	6690	**7520**	9484	11614
2524	3349	3978	4799	5186	6025	6757	7528	9602	**11728**
2538	3375	3996	4808	5191	6029	6770	7572	**9622**	12081
2614	3389	4027	**4817**	5203	6032	6771	7594	9625	12414
2649	3397	4042	4859	5209	6038	6844	7611	9636	12475
2656	3423	4057	4864	5222	6082	6851	7704	9655	14033
2709	3428	4119	4873	5235	6090	**6899**	**7711**	9658	14105
2723	3446	4234	4891	5241	6226	6933	7731	9698	14220
2731	3489	4236	4904	5272	6246	6967	8051	9732	14224
2744	3496	4304	4946	5292	**6262**	6973	8126	**9748**	15227
2802	3511	4310	4955	5297	6278	7062	8219	9910	
2839	3514	4317	4973	5342	6297	7108	8570	9983	
2904	3520	4319	4978	5373	**6335**	7117	8587	**10045**	
2946	3538	**4380**	4986	5395	6356	7171	8604	10061	

* Bolded masses match with the LC/ESI data.

Table 3.10 (M+H)$^+$ from *C. freundii* extract by LC/off-line MALDI.*

2064	2744	3463	4476	5857	**8278**	10649
2210	2755	3467	**4504**	**5936**	8318	**10681**
2215	2770	3481	4568	5995	8487	10948
2221	**2793**	3505	4681	6119	8699	11416
2238	2807	3531	4762	6185	8782	11424
2252	2831	3546	4977	6205	8802	**11748**
2319	2849	**3564**	4998	6256	8843	11761
2335	**2853**	**3581**	5090	6300	**9197**	**11870**
2352	2879	3669	5118	6547	9285	11902
2391	2958	3748	5171	6552	9353	12069
2431	2982	3923	5188	6666	9365	**12157**
2476	3029	3932	**5241**	6685	9368	12866
2500	3060	3951	**5256**	**6721**	**9522**	15519
2515	3065	3955	5299	6774	**9528**	15903
2520	3138	3971	5379	6792	9757	15920
2579	3176	3991	5458	6943	9867	15940
2584	3188	**4008**	5493	7023	**10108**	16774
2621	**3237**	4036	5549	**7335**	10121	**17177**
2636	3254	4069	5627	**7736**	10188	18552
2646	3290	4208	5646	**7752**	**10300**	
2662	3361	**4421**	5753	7866	10447	
2677	3370	**4435**	5792	7881	10522	
2689	**3449**	4454	5851	8226	10632	

* Bolded masses match with the LC/ESI data.

3.4 Discussion

3.4.1 Public Proteome Database

One of the critical issues in developing a strategy for bacterial identification based on searching a set of protein masses against protein mass databases is the selection and availability of databases. One route is to use the Internet-based public proteome database. Unfortunately, for a vast majority of bacteria, the number of entries in the current proteome database is still very limited, particularly in the low mass range (2-20 kDa), which is the main focus of sensitive bacterial identification by MS methods. This is true even for bacteria with complete genome databases. The lack of low mass protein information in the current proteome databases is attributed to the fact that it is difficult to accurately predict and detect small genes. The proteomes for some bacteria, such as *E. coli* and *B. subtilis*, are relatively well studied since they are often used as models to study gram-negative and gram-positive bacteria respectively. But it is still found that quite a high percentage of genes remains unassigned and a large number of open reading frames (ORFs) are poorly characterized for *E. coli*.[10, 11] According to NCBI (National Center for Biotechnology Information,), the genome of 50 bacteria has been completely sequenced (http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/eub_g.html, as of September, 2001). Among them, the number of protein entries in the corresponding proteome database is dramatically different. There are 2999 entries for *E. coli* between 2 and 20 kDa, 1516 entries for *B. subtilis*, whereas for some other species only very few entries exist. For instance, *Clostridium acetobutylicum* has only 40 entries and *Ureaplasma urealyticum* has only 30 entries. The dramatic differences in the number of protein entries are partially related to the different genome sizes, but, to most extend, is due to the poor characterization of gene products. Consequently, any attempt to identifying bacteria based on searching the current proteome database will inevitably be biased towards those bacterial species with relatively complete proteome databases.

The analysis of *B. megaterium* can be used as an example to demonstrate the limitation of using public proteome database for bacterial identification. Direct MALDI analysis showed 21 peaks (Table 3.6). Of those, only four matched with the *B. megaterium* proteome database (underlined). Note that there are only 55 protein entries in the mass range between 2 and 20 kDa in the proteome database of this bacterium.

Fifteen out of the 21 masses matched with the *E. coli* proteome database and nine masses matched with another *Bacillus* species, *B. subtilis*.

The LC/ESI-MS data from *B. megaterium* (Table 3.3) also produced protein masses that were better matched with those from other genera and species. Of the 47 protein components, only four matched with the *B. megaterium* database, while 19 masses matched with the *E. coli* database and 12 matched with *B. subtilis*. This example clearly demonstrates that the current public proteome databases are not sufficient for bacterial identification based on protein masses alone. It should be noted that one cannot use the normalized numbers of entries in these three databases for comparison, because the probability of detecting a given protein mass is not the same for all entries.

Besides the very limited proteome entries for most bacterial species, another major limitation of the current bacterial proteome database is that many of the protein masses listed in the database were derived from their genome sequences and were not confirmed experimentally. In reality, it is difficult to translate genome sequence or gene information into *in vivo* protein masses. The exact starting and ending sequences of a gene that will be used to express the protein can sometimes be difficult to predict. Post-translational modifications can change the molecular mass of a protein. Even though the genome database for many organisms is expanding rapidly, which will greatly facilitate the establishment of the genome-derived *proteome sequence* database, the establishment of the *proteome mass* database for the organism will always lag behind. Moreover, *in vitro* processes, such as protein modification or fragmentation during protein extraction or during MALDI and ESI sample preparation, can also alter the protein apparent masses. This issue will be discussed in Section 3.4.2.

3.4.2 Comparison of Three Data Sets

Instead of relying on the proteome database, one can establish protein mass databases using MS techniques. Different MS techniques were studied to evaluate the possibility of using them for bacterial protein mass database creation.

For *E. coli*, *B. megaterium* and *C. freundii*, direct MALDI analysis shows 29, 21, and 20 components. LC/ESI-MS shows 44, 47, and 46 components and LC/off-line MALDI shows 439, 286, and 157. Based on the different ionization methods and sample introduction procedures, it appears very likely that the ion suppression effects as well as variation in sample concentration have lead to the observed disparity in the number of detected species. Most chromatographic peaks in LC/ESI-MS exhibited mass spectral peaks from at most two or three components. In contrast five to ten components are detected in many fractions in LC/off-line MALDI. Note that each fraction used for the off-line MALDI analysis was concentrated by a factor of 50 as described in Section 3.2.3.2 before mixing with the second layer of matrix in a two-layer sample preparation.[12]

Although direct MALDI analysis is the most straightforward method of the three studied, it clearly suffers from the inability to detect a large number of components that are present in the cell extracts. This is not surprising given that many components in different concentrations are analyzed simultaneously. With on-line HPLC separation, ESI MS partially overcomes the problem of simultaneous detection of large numbers of species. However, it is clear that ion suppression still takes place within a mixture. Once a separation step has been included, the advantage that MALDI possesses over ESI for analysis of complex mixtures becomes evident. LC/off-line MALDI analysis remains slower and more labor intensive than online LC/ESI, yet it produces many more masses for possible inclusion in the proposed database.

It should be noted that not all of the masses detected from direct MALDI are present in the mass tables produced by LC/off-line MALDI. In one-dimensional HPLC separation, multiple components usually co-elute into the same fraction. Thus ion suppression remains an obstacle, albeit not to the same extent as with direct MALDI. The potential for protein loss during sample work-up also remains an issue. To mitigate these effects, sample preparation under different experimental conditions can be very helpful. Variations in extraction solvent and cell lysis method will undoubtedly result in the extraction of a different set of proteins or protein concentrations that may result in changes in the observed suppression effects. Work in this direction is currently underway in our laboratory.

Although LC/off-line MALDI detects more peaks than LC/ESI-MS, the masses detected by LC/ESI-MS are not necessarily a subset of those produced by LC/off-line MALDI. For *E. coli*, LC/ESI-MS produces 36 matches with the LC/off-line MALDI table. *B. megaterium* results in 29 of the 47 components matching with the results found by LC/offline MALDI. *C. freundii* has 33 out of 46 match with the MALDI data.

When comparing the results from direct MALDI and LC/ESI-MS analysis, it is found that for *E. coli*, 13 out of 29 masses observed in direct MALDI were also detected in LC/ESI-MS. 13 out of 21 masses observed in direct MALDI were also detected in LC/ESI MS for *E. coli*, 9 out of 20 masses observed in direct MALDI were detected in LC/ESI MS for *C. freundii*. Again this is an indication of the different suppression processes in MALDI and ESI.

In summary, the difference in sensitivity and the extent of ion suppression from MALDI and ESI results in the observation of different sets of proteins. To generate a mass table better representing the protein components in the cell extracts, it is important to combine the masses detected by the two ionization techniques.

3.4.2 Comparison of MS Data with Published Proteome Database

The current *E. coli* proteome database is quite extensive. However, Table 3.2, 3.5, and 3.8 show that most of the masses observed by the three methods are not listed in the proteome database. This is not surprising. Despite the completeness of its genome, it has been found that a significant percentage of *E. coli* genes or ORFs remains either unassigned or poorly characterized. Moreover, more than 60% of *E. coli* proteins are found to be proteolytically processed.[13] A variety of post-translational modifications, including methylation, acetylation and carboxylation, may have occurred for the proteins observed. Many of the observed proteins are likely the products of *in vivo* fragments from larger proteins, such as the cleavage of N-teminal methionine (Met) and signal peptides. *In vitro* fragmentation occurring during sample preparation might also attributes to the discrepancy of the observed and predicted protein masses. In addition, the formation of disulfide bonds of some proteins should not be neglected. For example, the major protein identified as 50S ribosomal protein S20 has a theoretical molecular weight of 7871 Da, but the measured protein mass by LC off-line MALDI is 7867 Da. The 4 Da mass difference is due to the two disulfide bonds which results in the loss of four hydrogen atoms. The identification of this protein will be discussed in Chapter 6.

Although the development of an appropriate searching algorithm can take into account those well-known post-translational modification and fragmentation information of well-characterized bacteria, it is not generally applicable for most poorly studied species. Most importantly, the *in vitro* fragmentation and modification under different sample preparation conditions is usually not predictable, and these information will not likely be included into proteome database. The fact that it is very often for the masses detected by MS to be different from the genome-predicted protein ones presents a major limitation for bacterial identification based on public proteome database using protein mass data alone.

3.4.3 Evaluation of the Applicability of Protein Mass Tables for Bacterial identification

One of the objectives in this work is to evaluate whether mass tables created by LC/MALDI and LC/ESI are sufficient for bacterial identification. Using the mass tables (Table 3.2-3.4 and 3.8-3.10) as initial mass databases, we tried to evaluate if it is possible to differentiate bacterial species by matching their protein masses obtained by direct MALDI with those listed in the mass tables.

Figure 3.3 is the mass spectrum of an *E. coli* sample grown in-house. The bacteria were grown in LB broth at 37 °C with shaking. Cells were harvested at 36 h, washed with sterile water, and lyophilized and stored below 0 °C before extraction. The singly charged protein m/z values are listed in Table 3.11. Compared to the tables generated by LC/off-line MALDI (Table 3.8-3.10), twenty four masses (bolded) matched with those listed in the *E. coli* table, four matched with *B. megaterium* and seven matched with *C. freundii*. When the MALDI data listed in Table 3.11 are compared to the mass tables from LC/ESI (Table 3.2-3.4), 11 masses (underlined) matched *E. coli*, two masses matched *B. megaterium* and three masses matched *C. freunidii*. These results were very encouraging. A similar comparison was made for the *B. megaterium* sample analyzed by direct MALDI (Table 3.6). 16 out of the 21 peaks matched with those listed in the LC/MALDI table of *B. megaterium*, 11 masses matched *E. coli*, and two matched *C. freundii*.

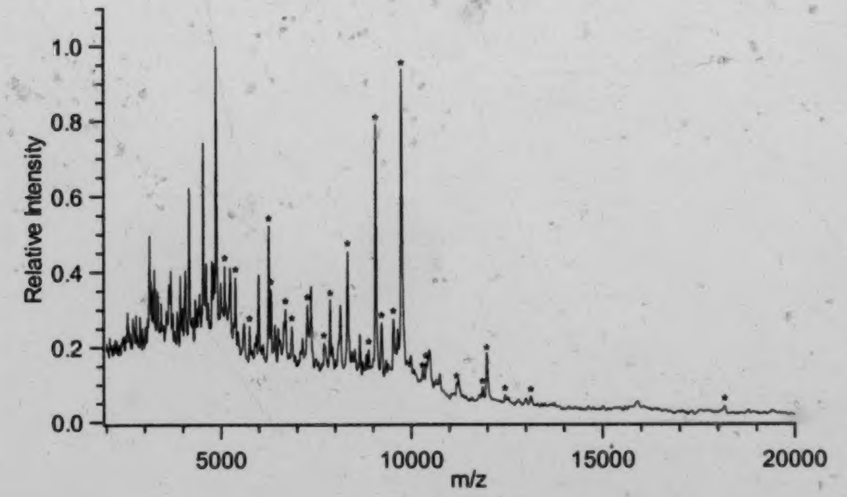

Figure 3.3 MALDI spectrum of *E. coli* harvested at 36 h. The peaks with * matched the masses in LC/off-line MALDI data (Table 3.8).

Table 3.11 $(M+H)^+$ of *E. coli* harvested at 36 h by direct MALDI (see text).

5097	**6697**	8139	**9741**	**11865**
5242	**6864**	**8327**	9988	**11977**
5381	7145	8649	**10300**	**12451**
5755	**7272**	**8884**	**10386**	13010
6255	7369	**9065**	10496	**13127**
6317	**7708**	**9225**	10750	15906
6420	**7868**	**9536**	**11186**	**18163**
6511	8096	9642	11229	

The results seem to be more ambiguous for identification. However, if LC/ESI results are to be used as the mass databases, 15 of the masses detected by direct MALDI matched *B. megaterium*, two matched *E. coli*, and none of these masses matched with any *C. freundii* proteins. Thus the LC/ESI-MS data give positive identification with high confidence. This result is not surprising. In direct MALDI analysis, only those proteins with relatively high abundance and those easily ionized are detected. The same is true for LC/ESI analysis, except that slightly different sets of proteins are detected due to the different ion suppression effect for the two techniques. For LC/MALDI analysis, however, low abundant protein species are very often detected due to pre-fractionation of the crude extracts as well as the post chromatographic concentration steps. The ability of detecting large sets of protein masses by LC/off-line MALDI analysis on one hand is very valuable as for the generation of a relatively complete protein mass table. On the other hand, it could result in congested protein mass information in the relative small mass window. The problem posed by congested masses information becomes prominent when the protein masses are not detected with high accuracy. As a result, it is relatively easy to give false positive identification.

One way to reduce the mass congestion is to increase the mass detection accuracy. With the most recent generation of high-resolution MALDI-TOF instruments, mass accuracy of better than 100 ppm can be obtained for analyzing proteins with masses of up to 30 kDa. Another way to reduce the unnecessary mass congestion is to exclude those low intensity protein peaks from mass tables by setting a reasonable signal intensity threshold for LC/off-line MALDI analysis. Those low intensity peaks often have poorly resolved protein masses. In addition, they are not likely to be detected by either direct

MALDI or LC/ESI analysis under the same sample preparation condition. These strategies will be considered in our future work to establish a much larger database.

3.4.4 Mass Database Creation

As discussed in the above sections, to create a valid protein mass database, it is important to combine the masses from the two different ionization techniques. In addition, different sample preparation methods have to be examined. Another important fact must be considered for bacterial protein mass database creation is that protein expression is very sensitive to the growth condition. Previous work has showed a mass table generated by LC/MALDI on a different batch of *E. coli* sample.[14] 169 out of 307 protein masses in that table matched with that in Table 3.8. The large numbers of different protein masses from the different batch sample suggests that bacteria grown under various conditions and harvested at different growth times need to be examined in order to create a valid mass table for each bacterium. Those protein masses consistently observed under each growth condition should be included into the mass table. This would compensate for the differences of protein expression under different environmental conditions. Even though different sets of protein masses may be detected in samples from different sources or from sample analyzed by different analytical methods, they can still be matched to the corresponding bacterium since the sets of masses should always reflect the bacterial proteome. In addition, a valid mass database should avoid any possible operator bias toward the protein mass results. This can be done by comparing the data obtained under nominally the same experimental condition from at least two laboratories.

3.5 Conclusions

We have shown that the present public proteome database has limited use for bacterial identification based on protein masses alone. Protein mass tables generated by LC/ESI-MS, direct MALDI and LC/off-line MALDI can potentially be used as a database for protein mass comparison. There are several merits of creating a database from mass spectrometric methods for bacterial identification. First of all, the proteome mass database is to be established for a single purpose, i.e., bacterial identification by mass spectrometry. Thus the initial database can be composed of a number of mass tables with each containing the consistently MS detectable proteins from an individual bacterium. The mass table can be rapidly created by using mass spectrometric techniques in conjunction with different sample preparation and handing methods. Secondly, the MS protocol used to create mass tables can be readily adapted and used to rapidly generate a data set for a new stain of bacteria. This is extremely important in order to keep the database current and meet special needs. Waiting for genome sequencing to be completed and then translating the genome into a proteome database is not practical in situations where new strains of bacteria are encountered. Thirdly, the mass database generated by MS can include many unique peptides or small proteins that are from the fragmentation of larger proteins either *in vivo* or *in vitro*. Many of these fragmented proteins are not included in the public proteome database. Fourthly, post-translational modified proteins or proteins being altered in masses during sample workup (e.g., oxidation) will be included in the MS generated database. Finally, the initial mass database can be readily expanded, if needed, to include other comparative parameters such as MS/MS spectra of intact proteins to increase the confidence level of identification. We note that, in creating the MS generated mass table, protein masses from the public database may be incorporated. If warranted, protein sequence information from the public database can also be used for data mining based on MS/MS of peptides.

3.6 Literatures Cited

1. Arnold, R.; Reily, J. *Rapid Commun. Mass Sepctrom.* **1998**, 12, 630.
2. Haag, A. M.; Taylor, S. N.; Johnston, K. H.; Cole, R. B. *J. Mass Spectrom.* **1998**, 33, 750.
3. Haddon, W. F.; Full, G.; Mandrell, R. E.; Wachtel, M. R.; Bates, A. H.; Harden, L. A. In *Proceedings of the 46th ASMS Conference on Mass Spectrometry and Allied Topics*, Orlando, FL, 1998; pp 177.
4. Nisson, C. L. *Rapid Commun. Mass Spectrom.* **1999**, 13, 1067.
5. Wang, Z.; Russon, L.; Li, L.; Roser, D.; Long, S. R. *Rapid Commun. Mass Spectrom.* **1998**, 12, 456.
6. Arnold, R.; Karty, J.; Ellington, A.; Reily, J. *Anal. Chem.* **1999**, 71, 1990.
7. Demirev, P. A.; Ho, Y. P.; Ryzhov, V.; Fenselau, C. *Anal. Chem.* **1999**, 71, 2732.
8. Whittal R. M.; Li, L. *Anal. Chem.* **1995**, 67, 1950.
9. Dunlop, K.Y.; Li, L. *J. Chromatogra A*, **2001**, 925, 123.
10. Blattner, F. R.; Plunkett, G.; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A.; Goeden, M. A., Rose, D. J.; Mau, B.; Shao, Y. *Science*, **1997**, 277, 1453.

11. Yamamoto, Y.; Aiba, H.; Baba, T.; Hayashi, K.; Inada, T.; Isono, K.; Itoh, T.; Kimura, S.; Kitagawa, M.; Makino, K.; Miki, T.; Mitsuhashi, N.; Mizobuchi, K.; Mori, H.; Nakade, S.; Nakamura, Y.; Nashimoto, H.; Oshima, T.; Oyama, S.; Saito, N.; Sampei, G.; Satoh, Y.; Sivasundaram, S.; Tagami, H.; Horiushi, T. *DNA Res*. **1997**, 4, 169.

12. Dai, Y.; Whittal, R. M.; Li, L. *Anal. Chem*. **1996**, 68, 2494.

13. Link, A. J.; Robison, K.; Church, G. M.; *Electrophoreis*, **1997**, 18, 1259.

14. Dai, Y.; Li, L.; Roser, D. C.; Long, S. R. *Rapid Commun. Mass Spectrom*. **1999**, 13, 73.

Chapter 4

Matrix-Assisted Laser Desorption Ionization Mass Spectrometry and Gel Electrophoresis Analysis of Bacterial Proteome from Rapid Solvent Extraction of Bacterial Cells [a]

4.1 Introduction

Matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (MS) has been used for the analysis of proteins from either cell lysates[1-12] or intact bacterial cells.[13-24] Recently, liquid chromatography/electrospray ionization-mass spectrometry (LC/ESI-MS) has also been applied for the analysis of low-mass bacterial proteins from cell extracts.[25-30] A subset of protein masses from a bacterial proteome is proposed to be used as the basis of bacterial discrimination. The confidence of bacterial identification will generally increase as the number of proteins detected increases. The number of proteins detected by MS clearly depends on how they are extracted from the bacterial cells. Many researchers have noted that bacterial proteins detected by MALDI or ESI from simple solvent extraction are usually low molecular mass species (MW<20,000 Da). A few reports showed weak MALDI signals at higher masses up to 70,000 Da[16,24] from whole cell analysis. Recently, Voorhees et al. showed some impressive results of high mass protein analysis from whole cell bacteria by MALDI.[24] They found that on-probe cell treatment with ethanol and MALDI sample preparation technique are critical in detecting proteins with MW>20,000 Da.[24]

---

[a] A form of this chapter is submitted as: Z. Wang, J. Zheng, L. Li, *"Matrix-Assisted Laser Desorption Ionization Mass Spectrometry and Gel Electrophoresis Analysis of Bacterial Proteome from Rapid Solvent Extraction of Bacterial Cells"*, *J. Mass Spectrometry*. Ms. Jing Zheng collected the in-gel digestion and peptide mass mapping data.

In this chapter, we report a study focusing on the issue of how rapid solvent extraction methods can affect MS detection of low and high mass proteins. Solvent extraction is widely used not only in MALDI[1-12] but also in ESI MS for rapid analysis of bacterial proteins.[25-27] In microbiology and other biochemical research, a number of cell lysing and protein extraction techniques can be used for different applications such as retention of protein bioactivity, reduction of proteolytic degradation, and elimination of contaminants that may influence further protein purification steps.[25] Most techniques are quite involved in terms of the apparatus required and experimental procedure. But for rapid MS analysis, a simple solvent system must be used for extraction in order to avoid any time consuming or laborious sample workup necessary for subsequent MS analysis. The most commonly practiced method is to suspend the bacteria cells in 0.1% trifluoroacetic acid (TFA), followed by vortexing and centrifugation.[3, 4] The use of other solvent systems have been studied and the differences in mass detection were noted in several reports.[2,6,12] However, to our knowledge, there is no systematic investigation into the effect of simple extraction methods on mass analysis that is specifically tailored to rapid bacterial identification by MS.

In this work, we used sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) to provide quantitative information on the relative abundance of extracted proteins. The results from SDS-PAGE are compared to the MALDI data of the same extracts. We demonstrated that both extraction and ion suppression play major roles in the outcome of mass analysis. Using peptide mass mapping and MS/MS database searching, we showed that the proteins extracted from simple solvent extraction methods are from bacterial cells, not from the contaminants associated with cell growth media or sample workup.

4.2 Experimental

4.2.1 Materials

Bacterial cells used in this work were from the Edgewood RDE Center at Aberdeen Proving Ground, MD. The cell growth condition was the same as described in Chapter 2. Spectrophotometric grade trifluoroacetic acid (TFA) was purchased from Sigma Aldrich Canada (Oakville, ON). HPLC grade acetonitrile and glacial acetic acid were from Fisher Scientific Canada (Edmonton, AB). Water was obtained from a Milli-Q Plus purification system (Millipore Corporation, Bedford, MA, USA).

4.2.2 Bacterial Protein Extraction

Bacterial proteins were extracted by solvent suspension as described previously.[6,7] About 1 mg of lyophilized bacterial sample was suspended in the extraction solvent (0.1% TFA, 40 mM Tris base or 50 mM NH_4HCO_3). The cell suspension was vortexed for 2 minutes and centrifuged at 14000 rpm. The supernatant was transferred to a vial. This process was repeated twice to maximize the extraction efficiency. Salts and small molecules were removed from the combined supernatants with Microcon-3 3000 Da molecular weight cutoff filters (Millipore Corporation, Bedford, MA, USA) before gel electrophoresis. For probe tip sonication, a Branson Sonifier (VWR, Bridgeport, NJ) was used and the cell suspensions were sonicated for 30s three times while cooling on dry ice, the cooling steps were used to avoid temperature increase. The duty cycle on the sonicator was set to 70%, and the output used was 3.

4.2.3 Total Protein Determination

The extracts were mixed with Coomassie plus protein assay reagent (Pierce, Rockford, IL). and the absorbance was read at 595 nm. A standard curve was prepared using bovine gamma globulin. The protein concentrations in the extracts were indicated in the Results and Discussion section.

4.2.4 SDS-PAGE

SDS-PAGE was carried out in a Bio-Rad mini-protein III system using 15% SDS polyacrylamide mini-gels. Prior to electrophoresis, protein samples were treated at ~95 C for 5 min in a pH 6.8 sample buffer containing 2% mercaptoethanol (v/v), 4% SDS, 12% glycerol, 50 mM Tris, and 0.01% bromophenol blue. Ten to 20 μL samples were loaded in the sample wells. The proteins were separated at constant current of 12 mA per gel for about one hour. Localization of the protein bands was carried out using Bio-Rad Biosafe Coomassie blue or silver staining kits.

4.2.5 Extraction of Intact Proteins from the Gel

Coomassie-stained gels were used for in-gel digestion and intact protein extraction. The protein bands of interest were cut out using a scalpel. Each gel piece was cut in half for extraction and put into a 0.6 mL microcentrifuge tube. Protein extraction was carried out by the addition of 5 to 20 μL of a saturated solution of α-cyano-4-hydroxycinnamic acid (HCCA) in 75% acetonitrile/ 0.1% TFA (aq). The gel was crushed in the eluting solvent using a 0.25 mL microcentrifuge tube. After vortexing for 30 s and centrifugation, the supernatant was collected for MALDI analysis.

4.2.6 In-Gel Digestion

Protein bands were excised from the gel and placed into 0.6-mL siliconized vials and rinsed with pH 8.5 100 mM NH_4HCO_3 buffer. The NH_4HCO_3 solution was removed by pipette and replaced with 10 to 20 μL of the same buffer solution with 10 ng/μL trypsin. The gels were crushed using thin 0.25 mL plastic vials and incubated for 2 to 5 hours at 37°C. Extraction of the peptides was accomplished with three 20 μL aliquots of

75% acetonitrile in 0.25 % TFA / water followed by 10 μL of acetonitrile. For each extraction step, the sample was vortexed for 20 s followed by 20 min of sonication. The pooled extracts were evaporated to dryness using a Speed-Vac.

4.2.7 Mass Spectrometry

MALDI MS was performed on a Voyager Elite MALDI MS instrument (PerSeptive Biosystems, Inc., Framingham, MA). A two-layer method was used for MALDI sample preparation.[9,10] One to 2 μL of the first-layer solution (10 mg of HCCA in 1 mL of 20% methanol/acetone (v/v)) was deposited onto the MALDI target, and evaporated to form a thin matrix layer. The second layer was prepared differently for analysis of proteins extracted from the gel and the crude bacterial cell extracts. For proteins extracted from gel, about 0.5 to 1 μL of gel protein extract in 50% acetonitrile or 40% methanol saturated with HCCA was deposited onto the first layer, and allowed to air dry. For the crude cell extracts analysis, about 1 μL cell extract was mixed with 5 to10 μL of the second layer matrix (HCCA or sinapinic acid (SA)), which was saturated in a mixture of formic acid/isopropanol/water (1/2/3, by volume). 1 μL of the second layer solution was deposited onto the first layer. With this preparation, about 0.2 μg of total proteins from cell extract was loaded onto the MALDI target. Mass calibration in MALDI was performed using external mass calibrants. Bradykinin and insulin chain B (oxidized) were used for calibrating the low mass region up to m/z 3500. The mass region from m/z 3500 to 20,000 was calibrated with insulin chain B, horse heart cytochrome c and its multiple charged species and dimer. The dimer of cytochrome c and the multiply charged species of BSA were used to calibrate the high mass range up to 67000 Da. The mass measurement accuracy is generally about 0.05%.

Nanospray ionization MS/MS was performed on an Esquire–LC ion trap mass spectrometer equipped with a NanoES interface (Agilent, Palo Alto, CA). Capillary LC MS/MS was performed on a Finnigan LCQ^{deca} system (Thermofinnigan, San Jose, CA) using data dependant analysis. The capillary column was 200 μm i. d. × 150 mm packed with 5 μm monomeric C_{18} media with 300 Å pore size (Vydac, Hespetia, CA). The mobile phases were 0.5% acetic acid in water (A) and acetonitrile (B). The gradient was 5-20% B in 10 min, 20-35% B in 30 min, 35-70% B in 10 min.

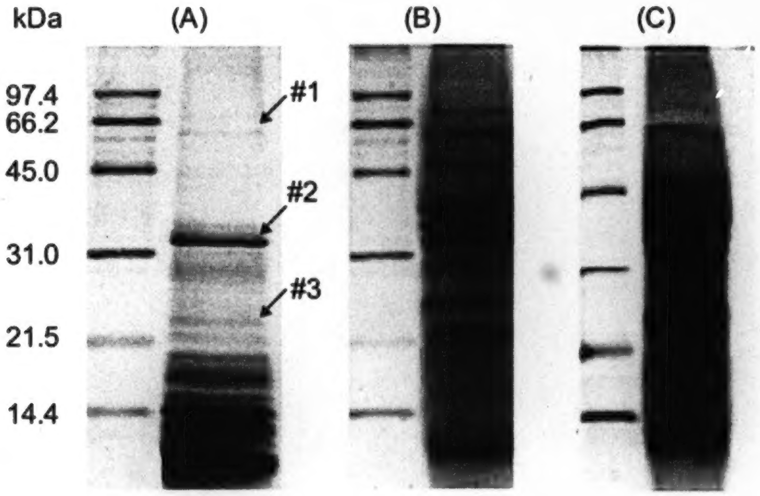

Figure 4.1 SDS PAGE of *E. coli* 9637 extracts prepared using different extraction solvents with vortexing: (A) 0.1% TFA, (B) 40 mM Tris-base (pH 9), and (C) 50 mM NH_4HCO_3 (pH 9). The proteins loaded into each lane were extracted from 0.2-0.5 mg starting lyophilized sample. Protein identification was carried out on the bands labeled with arrows (see text).

4.3 Results and Discussion

Figure 4.1 shows the gel images of proteins extracted from *E. coli* 9637 by simply vortexing the bacterial solvent suspension. Most protein components extracted by the 0.1% TFA aqueous solution have molecular weights below ~20 kDa (see Figure 4.1A). This is consistent with the observations from MALDI mass spectrometric analysis of the bacterial extract (see Figure 4.2A). The lack of mass spectral peaks from high mass proteins can be attributed to two possible reasons, namely suppression of the high mass ions by the easily ionizable low mass protein ions in MALDI and low efficiency for extracting high mass proteins. From the gel image shown in Figure 4.1A, it is clear that there are only very few visible bands in the high mass region. The relatively lower efficiency in extracting high mass proteins by 0.1% TFA with simple vortexing is therefore an important factor accounting for the absence of high mass ions in the mass spectra.

The type of solvent used for extraction is known to have a profound effect on extraction efficiency.[9] For example, it has been demonstrated that using different suspension solvents can result in the observation of different sets of low mass proteins. This observation was attributed to the different extents of cell lysis and/or the difference in the solubility of proteins in the suspension solvents. It has been shown that proteins with molecular masses up to 100 kDa can be extracted from bacteria cells in a Tris-base solution using vortexing and sonication.[4,5,19] Even without sonication, however, high mass proteins can still be extracted. This is shown in Figure 4.1B where the gel image was obtained from an *E. coli* extract prepared by using 40 mM Tris-base (pH 9) as extraction solvent with simple vortexing. Using Tris-base extraction, dark protein bands distributing in a much wider mass range are detected, compared to that obtained by TFA extraction (Figure 4.1A). In addition, more bands are detected in the mass range of 20 to 97 kDa than in the low mass region.

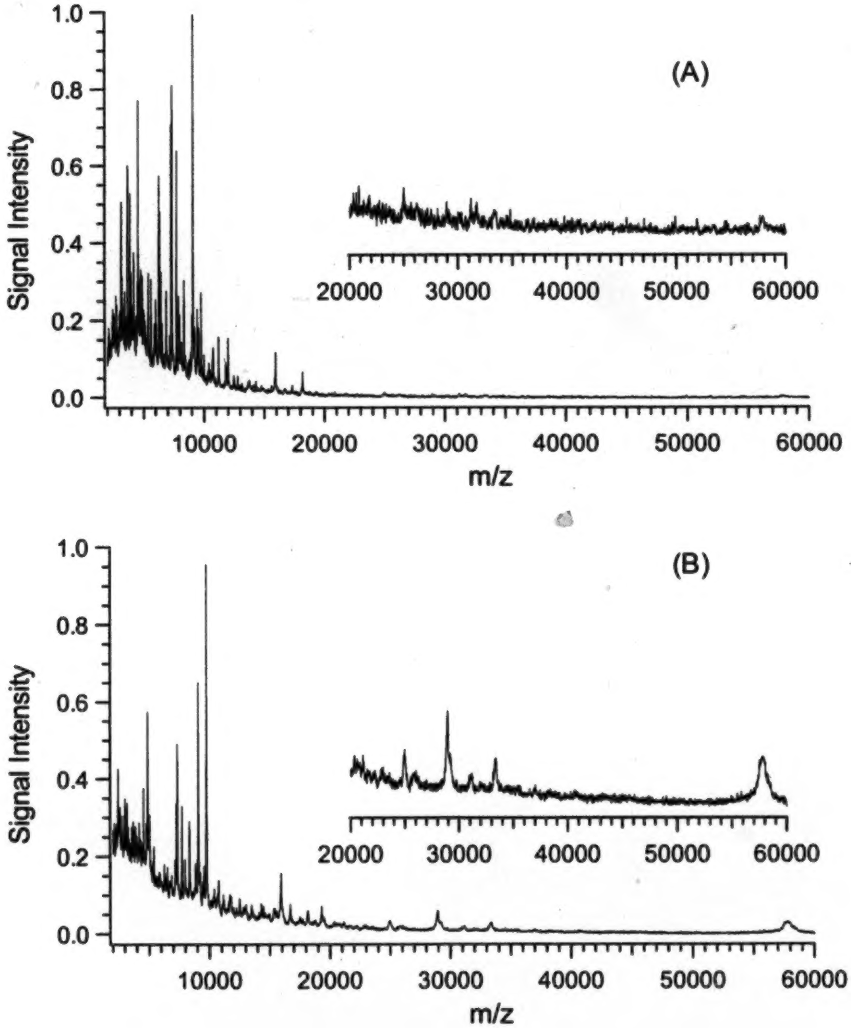

Figure 4.2 MALDI mass spectra of *E. coli* 9637 extracts prepared in different solvents with vortexing: (A) 0.1% TFA and (B) 40 mM Tris-base.

Using Tris-base at pH 9.0, the extraction should be more selective toward acidic proteins. The studies of low mass *E. coli* proteins[7,30,31] show that many proteins with molecular masses in the range of 2-20 kDa are the abundant, small alkaline proteins (high pI), such as ribosomal proteins and DNA binding proteins. Because of their basicity, they are readily extracted in 0.1% TFA, but not in the Tris-base buffer. This may explain why fewer proteins with molecular masses below 20 kDa were detected from *E. coli* sample with Tris-base as solvent. A 50 mM NH_4HCO_3 solution (pH 9) was also used to verify that pH is the key parameter in governing this differential extraction. The gel image from the NH_4HCO_3 extract is displayed in Figure 4.1C. The result is very similar to the Tris-base extract, indicating that pH is responsible for the extraction of different sets of proteins.

It is also possible that *E. coli* cells are lysed differently in the two different extraction solvents. From protein assay results (Table 4.1), the total protein content is much higher when Tris base is used as the extraction solvent. The improvement in cell lysis is likely another reason for the higher extraction efficiency of high mass proteins in Tris base. The effect of cell lysis on the display of protein contents will be discussed later.

Table 4.1 Protein assay results for extraction lyopholized *E. coli* sample using different solvent extraction methods.

	0.1% TFA		40 mM Tris base (pH 9)		50 mM NH_4HCO_3
	Vortexing	Sonication	Vortexing	Sonication	Vortexing
Total protein (μg)[a]	20	3.5×10^2	6×10^2	9×10^2	4×10^2

[a] Total protein from 3 mg starting lyopholized *E. coli* sample.

The differences between Tris-base and 0.1% TFA as extraction solvents were also evident in their MALDI spectra (see Figure 4.2). More peaks are detected in the low mass region from the TFA extract than that from the Tris-base extract. On the other hand, a few peaks in the 20-60 kDa mass range were detected from the Tris-base extract, but not from the TFA extract.

Although extraction of high mass proteins is not efficient using 0.1% TFA as suspension solvent, several bands above 20 kDa are still visible, as shown in Figure 4.1A. These proteins were not detected by direct MALDI analysis of the TFA extract, but could be detected after gel electrophoresis. As an example, Figure 4.3A shows molecular ion peaks around 58 kDa in the MALDI spectrum obtained by extracting the protein(s) from gel band #1 of Figure 4.1A. This result indicates that ion suppression prevents the detection of these high mass proteins during direct analysis of the cell extract.

Another important question is related to the origins of the low intensity bands shown in Figure 4.1A. They could either be from the bacteria cells or from background contaminants associated with the culture medium.[31] To find their source, we performed in-gel digestion followed by peptide mass mapping and MS/MS to identify some of these proteins. Figure 4.3B shows the MALDI spectrum of the tryptic digest from gel band #1 of Figure 4.1A. These tryptic peptide masses were used to search for the proteins in *E. coli* proteome database using the UCSF MS-Fit program. The proteins were identified as *E. coli* flagellins. There are 24 peptides matching with the theoretical digest of *E. coli* flagellins and the sequence coverage is 64%. To confirm the identity, the same tryptic digest was analyzed by nanospray ESI MS/MS in an ion trap mass spectrometer. Partial amino acid sequences of several peptides were obtained. One example is shown in Figure 4.3C for the MS/MS spectrum of a tryptic peptide with MH^+ at 632.4 Da. The sequence, LSSGLR, matches with *E. coli* flagellins. Other sequences found include

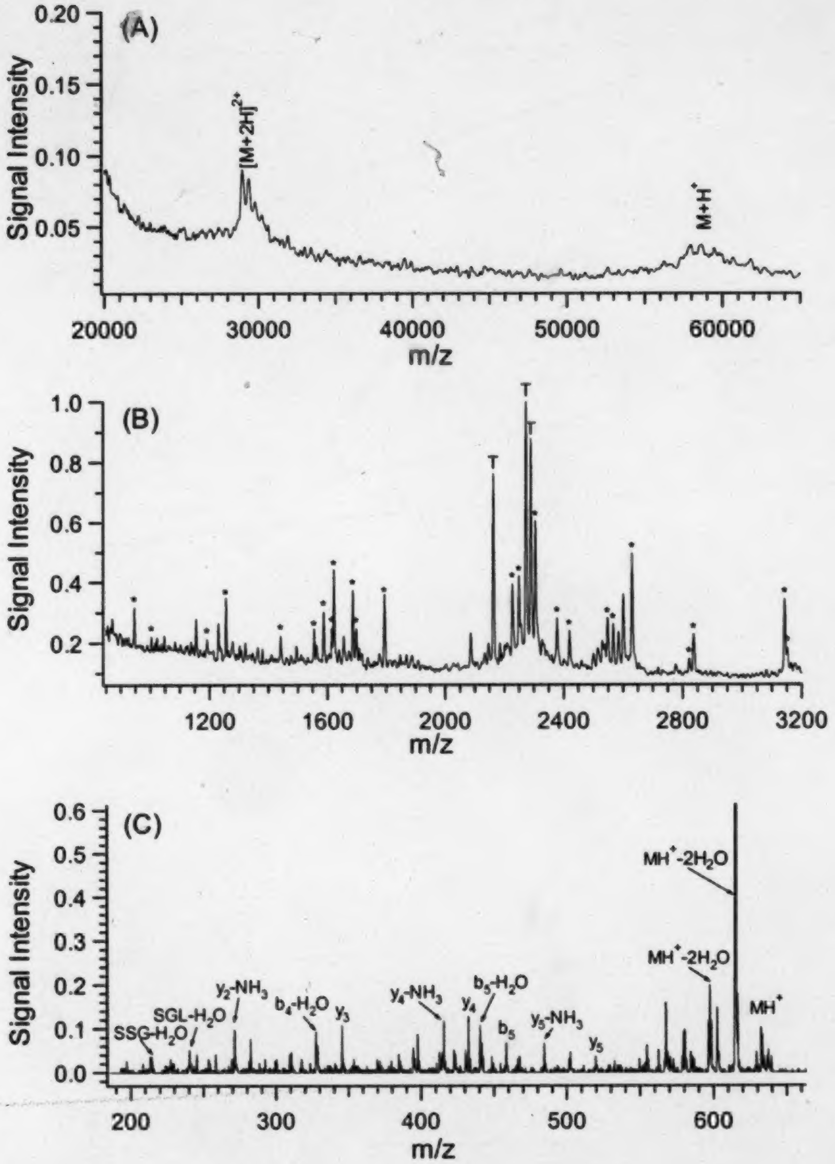

Figure 4.3 (A) MALDI mass spectrum of gel band #1 of Figure 4.1A after extraction. (B) MALDI mass spectrum of the in-gel digest of gel band #1. (C) ESI MS/MS spectrum of a tryptic peptide with MH^+ at 632.4 Da.

GLTQAAR (716.4 Da), LKDGDSVAVAAQK (1031.7 Da) and FTSNIK (709.4 Da). These sequences combined with the peptide mass map confirm that this protein is *E coli* Flagellins.

The major protein in band #3 of Figure 4.1A was identified as FKBP-type peptidedyl-polyl cis-trans isomerase FKPA precursor(ppiase)(Rotamase). The amino acid sequence of this protein is as follows,

MKSLFKVTLLATTMAVALHAPITFAAEAAPATAADSKAAFKNDDQKSAYALGASLGR
YMENSLKEQEKLGIKLDKDQLIAGVQDAFADKSKLSDQEIEQTLQAFEARVKSSAQAKM
EKDAADNEAKGKEYREKFAKEKGVKTSSTGLVYQVVEAGKGEAPKDSDTVVVNYKGT
LIDGKEFDNSYTRGEPLSFRLDGVIPGWTEGLKNIKKGGKIKLVIPPELAYGKAGVPGIPP
NSTLVFDVELLDVKPAPKADAKPEADAKAADSAKK

The protein has a molecular mass of 28882 Da, whereas the protein detected by MALDI from gel extract is about 26200 Da (spectrum not shown). This can be explained by the fact that this protein is localized in the periplasmic region, and a short piece of N-terminal signal peptide (amino acids 1-25, bold face) is cleaved from its mature form. Thus the theoretical molecular weight of 26224 Da closely matches the MALDI molecular weight result. The peptide mass mapping results are shown in Table 4.2 with 55% protein sequence coverage. A tryptic peptide with MH^+ at 2249.5 Da matching the underlined sequence shown above was detected from the tryptic digest, which further confirmed the loss of a short piece of signal peptide.

Table 4.2 Tryptic peptides matched protein FKBP-type peptidedyl-polyl cis-trans isomerase FKPA from gel band #3 [Figure 4.1A]

	Protonated Mass $[M+H]^+$	
Tryptic Fragments	Theoretical Fragments	Observed Fragments
84-91	818.5	817.9
95-102	833.4	832.9
95-104	1018.5	1018.3
23-33	1065.6	1065.4
192-202	1199.7	1199.8
84-94	1222.6	1222.7
190-202	1440.9	1441.0
133-147	1621.8	1621.9
229-245	1686.9	1687.2
86-102	1793.8	1793.8
49-65	1847.0	1846.8
68-83	1877.9	1878.4
23-40	1948.2*	1948.5*
1-22	**2249.5***	**2250.1***
115-137	2306.6*	2305.9*
118-147	3143.5*	3143.9*

* Average masses

Peptide mass mapping alone did not provide adequate information for positive protein identification from the dark band with apparent mass of about 32 kDa (band #2 of Figure 4.1A). This is probably due to the presence of multiple proteins in this band. Using capillary LC MS/MS, five proteins were identified and the results are listed in Table 4.3. Each of the proteins was identified based on the sequencing information from several peptides. Figure 4.4 shows an example MS/MS spectrum obtained by LC MS/MS. A doubly charged tryptic peptide with $[M+2H]^{2+}$ at m/z 883.8 was selected as the parent ion; the peptide sequence was identified as ALAINLVDPAAAGTVIEK from D-galactose binding periplasmic Protein. Excellent matching between the experimental

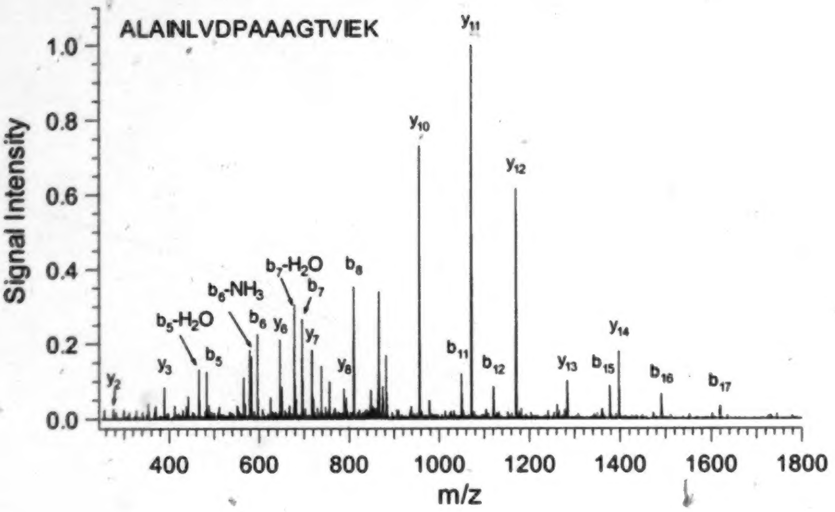

Figure 4.4 MS/MS spectrum of a doubly charged tryptic peptide with $[M+2H]^+$ at 883.8 Da from DGAL_ECOLI.

b and y peaks and the predicted fragment peaks (see Table 4.3) points to a high confidence of the identification of this peptide. Note that all of the proteins identified (see Table 4.3) are located in the periplasmic region with recognized N-terminal signal peptides. The matured proteins most likely have lost their signal peptides, resulting in molecular masses between 31-35 kDa, which is consistent with the location of the gel band.

The above results on positive identification of bacterial proteins in bands #1 to #3 (Figure 4.1A) suggest that some high mass proteins were extracted from *E coli* by using 0.1% TFA, albeit at lower efficiency compared to the low mass proteins.

Table 4.3 Proteins identified from gel band #2 of Figure 4.1A.

Protein	Identified Peptides (MH$^+$)	#b/y	Xcorr	dCn
DGAL_ECOLI (35712Da, P02927)	ALAINLVDPAAAGTVIEK (1765.8 Da)	14/12	6.15	0.66
	ALAINLVDPAAAGTVIEKAR (1993.1 Da)	12/10	5.72	0.59
	GQNVPVVFFNK (1248.5 Da)	7/9	3.26	0.54
	QNDQIDVLLAK (1256.7 Da)	8/9	3.74	0.57
	GQNVPVVFFNKEPSRK (1846.0 Da)	8/8	4.96	0.51
	VPYVGVDKDNLAEFSKK (1909.0 Da)	11/11	6.01	0.69
	SGALAGTVLNDANNQAK (1644.7 Da)	14/14	5.00	0.54
	ALDSYDKAYYVGTDSK (1796.6 Da)	12/12	5.80	0.70
YBEJ_ECOLI (33420Da, P37092)	DHGDSFRTLESGR (1477.6 Da)	9/8	2.30	0.40
	IISAKDHGDSFRTLESGR (1990.4 Da)	9/4	3.65	0.56
	ALFKEPNDKALN (1359.6 Da)	8/5	2.48	0.58
	GGDIKDFANLK (1178.6 Da)	8/7	2.73	0.30
	ESSVPFSYYDNQQK (1691.6Da)	11/10	4.77	0.44
	QAAFSDTIFVVGTR (1513.2 Da)	11/10	5.44	0.72
	KGGDIKDFANLK (1305.7 Da)	10/9	4.76	0.54
	VVGYSQDYSNAIVEAVK (1842.1 Da)	14/11	5.36	0.44
	NGVIVVGHR (950.7 Da)	6/8	3.78	0.53
	VVGYSQDYSNAIVEAVKK (1970.3 Da)	8/13	3.41	0.61
	LIPITSQNR (1041.8 Da)	5/6	2.78	0.47
ASG_ECOLI (36851Da, P00805)	TNTTDVATFK (1098.2 Da)	6/8	3.42	0.64
	SVNYGPLGYIHNGK (1519.0 Da)	10/8	4.90	0.25
	VGIVYNYANASDLPAK (1695.7 Da)	12/11	5.80	0.67
	HTSDTPFDVSKLNELPK (1929.1 Da)	11/13	5.20	0.58
	SVFDTLATAAK (1123.5 Da)	9/9	4.58	0.53
	VGVENLVNAVPQLK (1480.4 Da)	9/10	5.17	0.61
PSTS_ECOLI (37024Da, P06128)	TNIKDSSGKPLY (1323.4 Da)	8/10	3.33	0.44
	LISADGKPVSPTEENFANAAK (2160.1 Da)	16/16	6.30	0.64
	LPGAIGYVEYAYAK (1515 Da)	10/8	4.16	0.58
YDGH_ECOLI (33903Da, P76177)	AALAAGGEAAK (929.5 Da)	9/6	3.06	0.51
	FFETQSSK (974.2 Da)	6/7	3.09	0.50
	ITAFIYKK (983.5 Da)	6/6	2.28	0.19
	FNAIGEAVK (948.5 Da)	8/8	3.71	0.31
	AKGAYSFYIVR (1274.7 Da)	8/5	3.16	0.34
	KVEIPGVATTASPSSEVGR (1886.8 Da)	14/11	6.17	0.64
	GNNLTVSADLYK (1294.6 Da)	6/7	2.46	0.50

To find out if the lower efficiency in extracting high mass proteins using 0.1% TFA is a general trend, we tried several other bacteria. Figure 4.5 shows the gel images of protein extracts of *Citrobacter* freundii (*C. freundii), Aeromonas hydrophilia* (*A. hydrophilia*), and *Bacillus cereus* (*B. cereus*). More numerous and more intense bands with molecular masses below 30 kDa are displayed in all cases. We also tried to use 40 mM Tris base as extraction solvent (data not shown). Compared to the TFA extraction results, more proteins in the higher mass range were observed using Tris base. We note that in Figure 4.5 more high mass proteins were detected compared to the *E. coli* extract in 0.1% TFA. It is plausible that these bacterial cells were better lysed during sample manipulation prior to the extraction, or there are more basic proteins in the high mass range for these bacteria.

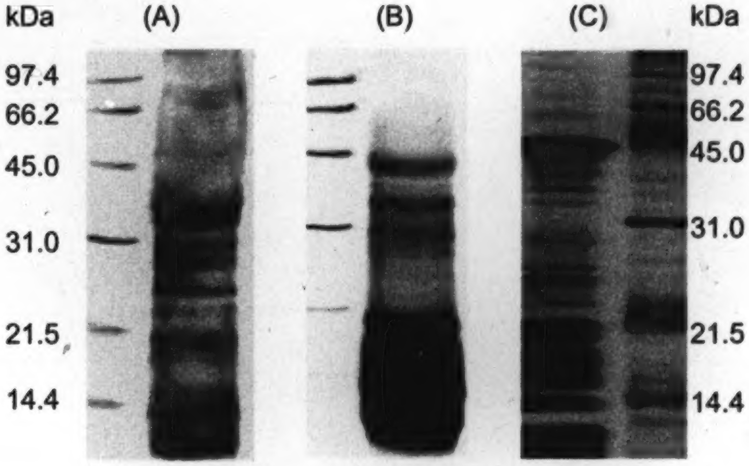

Figure 4.5 SDS PAGE image of proteins extracted from different bacteria with 0.1% TFA: (A) *C. freundii*, (B) *A. hydrophilia*, and (C) *Bacillus cereus*. The proteins loaded into each lane were extracted from 1 mg lyophilized bacterial samples.

Simple solvent extraction is sometimes combined with a physical cell lysis technique to assist cell lysis and protein extraction. Low extraction efficiency of high mass proteins could be caused by insufficient cell breakage. From earlier work on *E. coli*,[7, 30] proteins inside the cytoplamic membrane as well as proteins in the periplasmic region were identified. Consequently, it is believed that the bacterial cells were at least partially broken during bacterial manipulation and MALDI sample preparation. The high molecular mass flagellins identified in this work are a group of proteins located outside the cell wall. Thus it is not surprising that they are easily extracted with only vortexing.

For effective releasing of the proteins inside the cell wall, the bacterial cells have to be lysed more completely. This is especially important for higher molecular mass proteins, since they tend to be trapped in the partially broken cell debris. It is known that the disruption of bacterial cells by physical as well as chemical manipulation can help the recovery of higher mass proteins.[28] A seemingly effective method to assist the cell lysis during simple solvent extraction is the use of probe tip sonication. We have examined how this method affects the MS detection of low and high mass proteins.

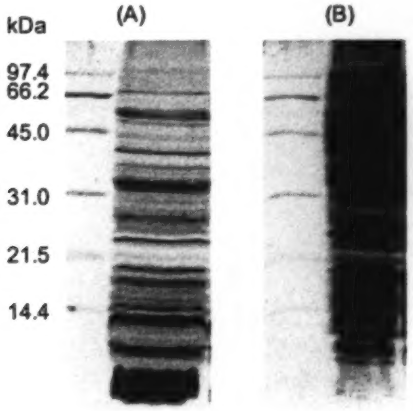

Figure 4.6 SDS PAGE image of *E. coli* proteins using probe tip sonication in different extraction solvents: (A) 0.1% TFA and (B) 40 mM Tris-base. The proteins loaded into each lane were extracted from 0.2-0.5 mg lyophilized cells.

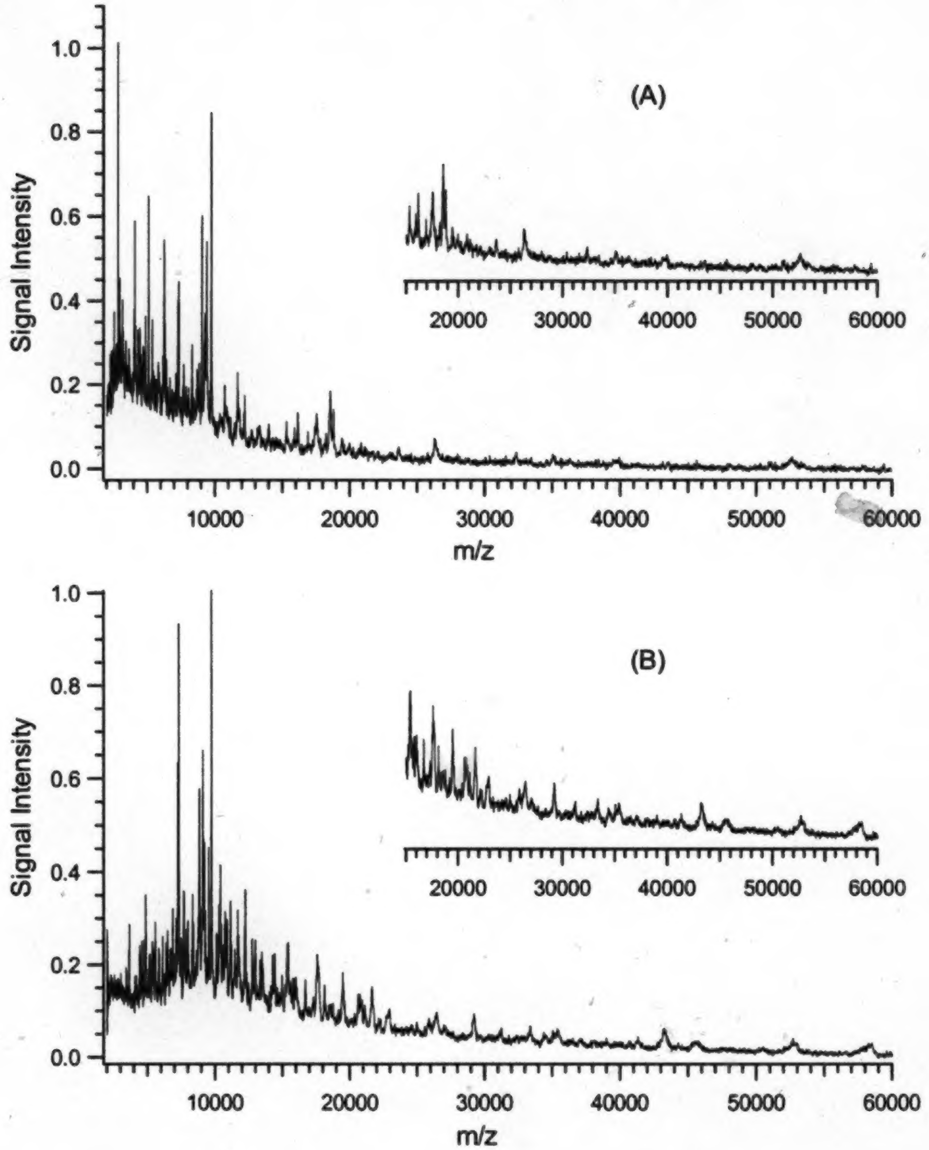

Figure 4.7 MALDI analyis of *E. coli* extract prepared by probe tip sonication in 40 mM Tris base (A) spectrum of crude extract, (B) spectrum of extract after 10 kDa molecular mass cutoff.

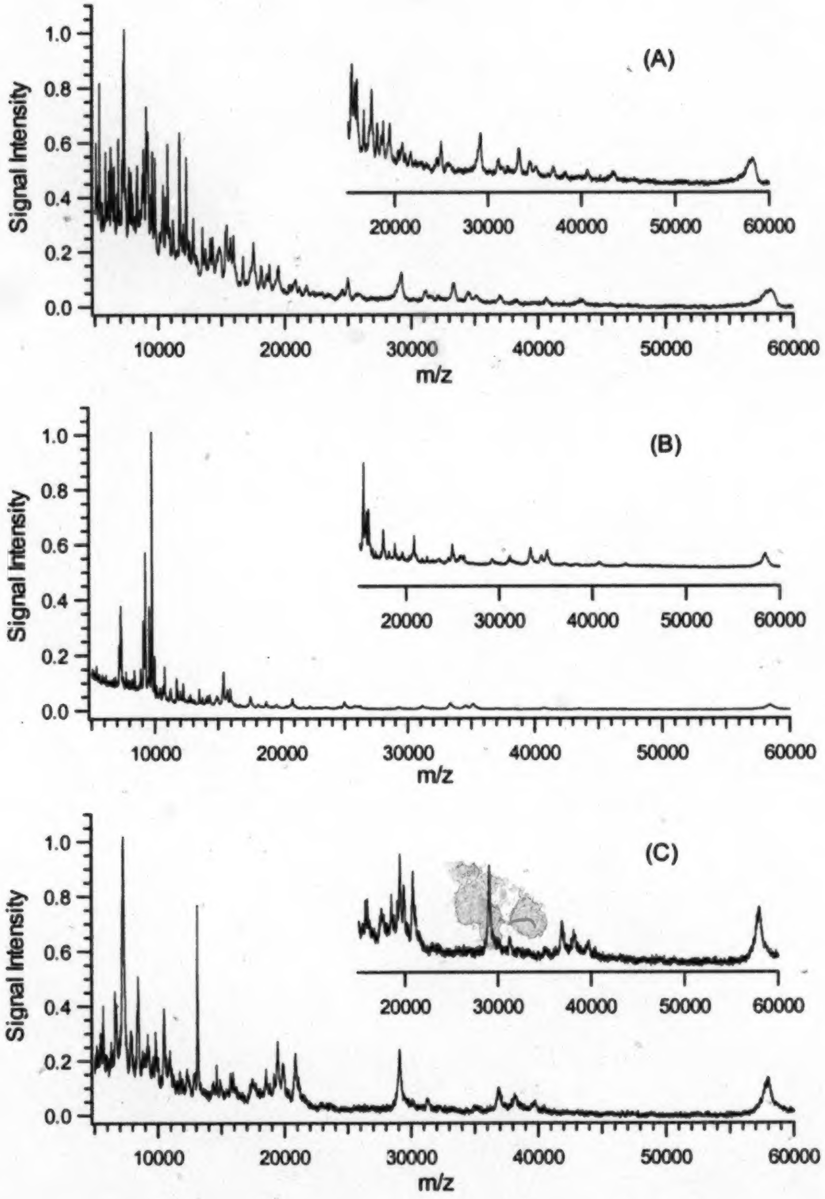

Figure 4.8 MALDI spectra of *E . coli* extract in 0.1% TFA using probe tip sonication. (A) HCCA as matrix using two-layer MALDI sample preparation, (B) SA as matrix using two-layer MALDI sample preparation, (C) HABA as matrix using dried droplet sample preparation.

The protein assay results (Table 4.1) of *E. coli* extracts clearly shows that probe tip sonication dramatically increases the total protein extracted, especially when 0.1% TFA was used as solvent. The combination of sonication with the use of basic solvent results in the extraction of the largest amount of proteins. Figure 4.6 shows the gel images of *E. coli* proteins extracted using different solvents by probe tip sonication. It is clear that many more proteins were extracted with sonication instead of simple vortexing using either 0.1% TFA or 40 mM Tris base solution. For the 0.1% TFA extract, more protein bands in the high molecular mass range were detected using sonication instead of simple vortexing, which confirmed that poor cell breakage is one of the main reasons for the low efficiency in extracting high mass proteins. It should be noted that, in the case of TFA extraction, low mass proteins are still the dominant components in the extract even when sonication is used to assist the cell lysis. This result again indicates the general low solubility of high mass proteins in this solvent.

Direct MS analysis of cell extracts prepared by probe tip sonication did not necessarily give higher quality MALDI spectra compared to that obtained by simple vortexing. In Figure 4.7A, fewer protein peaks was detected compared to that shown in Figure 4.2B, even though it is clear that more proteins were in the extracts from both gel electrophoresis analysis and protein assay. This is most likely caused by a larger suppression effect due to the presence of a large number and amount of proteins. Dilution of the mixture did not result in the increase in the number of proteins detected. Note that sample dilution does not change the relative amounts of proteins present in the mixture. Ion suppression is likely due to the presence of too many proteins with similar amounts. If HPLC fractionation or other separation techniques is used to separate the complex mixture into fractions containing a few number of proteins, followed by the analysis of individual fractions, a greater number of proteins are expected to be detected from the extracts. This has been illustrated in Chapter 3. Figure 4.7B shows the MALDI spectrum of the same extract after 10 kDa molecular weight cutoff by Microcon-10 filter, which is a simple method to filter out low mass proteins as well as salts in the extracts. It is not surprising to observe the enhancement in the detection of high mass proteins.

It should be noted that the matrix sample preparation method has a significant effect on high mass detection. For bacterial cell extract analysis, we prefer to using HCCA as matrix prepared by the two-layer method, since it is generally more effective in handling protein samples containing salts and shows less mass discrimination for protein mixture analysis.[32] Figure 4.8 illustrates the performance of this method compared to other commonly used sample preparation protocols in handling high mass proteins. A larger number of proteins distributed from 2000 to 60000 Da was detected in Figure 4.8A when a two-layer method was used to prepare the MALDI sample with HCCA as matrix. In the case of using SA as matrix (Figure 4.8B), a number of high mass proteins are detected while fewer low mass peaks were observed compared with that using HCCA as matrix. This is consistent with the notion that SA gives better detection for high mass proteins. HABA was also tested. The MALDI result is shown in Figure 4.8C. The dried droplet method was used to prepare the MALDI sample for HABA due to its high solubility in water and the difficulty in performing a two-layer sample preparation. Table 4.4 lists the m/z of proteins detected using the three methods. The protein detected from *E. coli* extract prepared by simple vortexing in 0.1% TFA was also listed in Table 4.4. As Table 4.4 shows, most protein masses detected by the SA and HABA matrix preparation methods are detected by the HCCA method. Compared to the extract prepared by simple vortexing, more components were detected when proteins were extracted by probe tip sonication. When the sets of protein masses are used to search bacterial protein mass databases, the detection of a greater number of proteins in a wider mass window by using an optimized sample extraction and MALDI preparation method will certainly result in better discrimination among the possible bacteria candidates.

4.4 Conclusions

We have shown that the method of solvent extraction has a profound impact on direct MALDI analysis of bacterial proteomes. Also under proper conditions, high mass proteins with molecular weights of up to 60,000 Da can be readily detected by MALDI TOFMS. When a 0.1% TFA aqueous solution is used as suspension solvent in combination with either simple vortexing or probe tip sonication, the extracted proteins are mainly from the low mass region (MW below 20 kDa). In contrast, when 40 mM Tris-base is used as suspension solvent, higher molecular weight proteins are efficiently extracted, particularly in the 20-100 kDa mass range. The use of probe tip sonication to lyse the bacterial cells results in more complete cell breakage, and consequently results in the releasing of more higher molecular mass proteins. However, the extraction efficiency and ion suppression effect need to be balanced to arrive at an optimal detection of the bacterial proteome by direct MALDI. It should be noted that, in the application of the MALDI MS technique to real world samples, other issues in dealing with other hard-to-break sporulated bacteria[10] and/or bacterial mixtures still need to be addressed in the context of optimal extraction and reduced ion suppression.

Table 4.4 Protein MH^+ detected from 0.1% TFA extracts (sonication or vortexing) using different MALDI matrices and sample preparation methods (see text), those marked with * match the results shown in the first column.

HCCA (sonication)	SA (sonication)	HABA (sonication)	HCCA (vortexing)
4071	4612	5684	4771*
4366	4771*	6600	5096*
4772	5381*	7273*	5381*
5097	7181	7868*	6055*
5380	7273*	8438	6256*
6256	7334*	9191*	6316*
6316	7707*	10476	6414
6395	8849*	13095	6512
6508	9066*	14608*	6860*
6858	9122*	15913*	7143*
7142	9226*	17512*	7272*
7179	9536*	19426	7333*
7272	9740*	19872	7707*
7333	9952	20824*	7868*
7661	10113	36960*	8326*
7707	10601*	37279	8994
7868	10750*	38150*	9065*
8328	11216*	39742	9191*
8848	11688*	58160*	9225*
8893	11975*		9431
9065	12220*		9536*
9120	12653		9740*
9192	12768*		9982
9226	13480*		10245
9535	13737*		10386
9685	14095*		10466
9739	14281*		10651
10373	14870		10694
10747	15411*		10733
10600	15694*		11185*
11116	15920*		11781
11185	17515*		11866*
11689	18768*		11976*
11866	20859		12768*
11976	21717*		14285*
12214	22250		15915*
12474	24968*		18160*
12768	26247		
13482	31132*		
13736	33340*		
14100	34520*		

Table 4.4 Continued

HCCA (sonication)	SA (sonication)	HABA (sonication)	HCCA (vortexing)
14282	35164*		
14605	40673*		
15412	58222*		
15693			
15920			
16689			
17513			
18164			
18508			
18769			
20819			
21721			
24969			
31148			
33323			
34511			
35172			
36953			
38158			
40669			
43400			
58190			

4.5 Literatures Cited

1. Cain, T. C.; Lubman, D. M.; Weber, Jr., W. J.; *Rapid Commun. Mass Spectrom.* **1994**, 8, 1026.

2. Krishnamurthy, T.; Ross, P. L.; Rajamani, U. *Rapid Commun. Mass Spectrom.* **1996,** 10, 883.

3. Chong, B. E.; Wall, D. B.; Lubman, D. M.; Flynn, S. J. *Rapid Commun. Mass Spectrom.* **1997,** 11, 1900.

4. Van Adrichem, J. H. M.; Bornsmen, K. O.; Conzelmann, H.; Gass, M. A. S.; Eppenberger, H.; Kresbach, G. M.; Ehrat, M.; Leist, C. H. *Anal. Chem.* **1998,** 70, 923.

5. Easterling, M. L.; Colangelo, C. M.; Scott, R. A.; Ameter, I. J. *Anal. Chem.* **1998,** 71, 3226.

6. Wang, Z.; Russon, L. M.; Li, L.; Roser, D. C.; Long, S. R. *Rapid Commun. Mass Spectrom.* **1998,** 12, 456.

7. Dai, Y.; Li, L.; Roser, D. C.; Long, S. R. *Rapid Commun. Mass Spectrom.* **1999,** 13, 73.

8. Holland, R. D.; Duffy, C. R.; Rafii, F.; Sutherland, J. B.; Heinze, T. M.; Holder, C. L.; Voorhees, K. J.; Lay, Jr., J. O. *Anal. Chem.* **1999,** 71, 3226.

9. Arnold, R. J.; Reilly, J. P. *Anal. Biochem.* **1999,** 269, 105.

10. Birmingham, J.; Demirev, P.; Ho, Y.; Thomas, J.; Bryden, W.; Fenselau, C. *Rapid Commun. Mass Spectrom.* **1999,** 13, 604.

11. Wall, D. B.; Lubman, D. M.; Flynn, S. J. *Anal. Chem.* **1999,** 71, 3894.

12. Domin, M. A.; Welham, K. J.; Ashton, D. S. *Rapid Commun. Mass Spectrom.* **1999,** 13, 222.

13. Holland, R. D.; Wilkes, J. G.; Rafii, F.; Sutherland, J. B.; Persons, C. C.; Voorhees, K. J.; Lay, Jr., J. O. *Rapid Commun. Mass Spectrom.* **1996,** 10, 1227.

14. Krishnamurthy, T.; Ross, P. L. *Rapid Commun. Mass Spectrom.* **1996,** 10, 1992.

15. Claydon, M. A.; Davey, S. N.; Edwards-Jones, V.; Gordon, D. B. *Nature Biotechnol.* **1996,** 14, 1584.

16. Welham, K. J.; Domin, M. A.; Scannell, D. E.; Cohen, E.; Ashton, D. S. *Rapid Commun. Mass Spectrom.* **1998**, 12, 176.

17. Haag, A. M.; Taylor, S. N.; Johnston, K. H.; Cole, R. B. *J. Mass Spectrom.* **1998,** 33, 750.

18. Arnold, R. J.; Karty, J. A.; Ellington, A. D.; Reilly, J. P.; *Anal. Chem.* **1999**, 71, 1990.

19. Saenz, A. J.; Pertersen, C. E.; Valentine, N. B.; Gantt, S. L.; Jarman, K. H.; Kingsley, M. T.; Wahl, K. L. *Rapid Commun. Mass Spectrom.* **1999,** 13, 1580.

20. Lynn, E. C.; Chung, M.; Tsai, W.; Han, C. *Rapid Commun. Mass Spectrom.* **1999,** 13, 2022.

21. Leenders, F.; Stein, T. H.; Kablitz, B.; Franke, P.; Vater, J. *Rapid Commun. Mass Spectrom.* **1999**, 13, 943.

22. Winkler, M. A.; Uher, J.; Cepa, S. *Anal. Chem.* **1999**, 71, 3416.

23. Evason, D. J.; Claydon, M. A.; Gordon, D. B. *Rapid Commun. Mass Spectrom.* **2000**, 14, 669.

24. Madonna, A. J.; Basile, F.; Ferrer, I.; Meetani, M. A.; Rees, J. C.; Voorhees, K. J. *Rapid Commun. Mass Spectrom.* **2000,** 14, 2220.

25. MacNair, J. E.; Opiteck, G. J.; Jorgenson, J. W.; Moseley, A. M. *Rapid Commun. Mass Spectrom.* **1997**, 11, 1279.

26. Krishnamurthy,T.; Davis, M. T.; Stahl, D. C.; Lee, T. D. *Rapid Commun. Mass Spectrom.* **1999**, 13, 39.

27. Dalluge, J. J.; Reddy, P. *Biotechniques* **2000**, 28, 156.

28. Cull, M.; McHenry, C. S.; In *Guide to Protein Purification,* Deutscher MP (ed), Academic Press: New York 1991, 147.

29. Molloy, M. P.; Herbert, B. R.; Walsh, B. J.; Tyler, M. I.; Traini, M.; Sanchez, J.; Hochstrasser, D. F.; Willams, K. L.; Gooley, A. A. *Electrophoresis* **1998**, 19, 837.

30. Wang, Z.; Doucette, A.; Li, L. in preparation.

31. Lay, Jr., J. O. *Trends in Anal. Chem.* **2000**, 19, 507.

32. Dai, Y.; Whittal, R. M.; Li, L. *Anal. Chem*, **1999**, 71, 1087.

Chapter 5

Identification of Low Mass Bacterial Proteins and Their Post-translational Modifications by HPLC Separation, Enzymatic Digestion and MALDI

5.1 Introduction

In the previous chapters, it has been shown that low mass bacterial proteomes can be analyzed by different MS techniques. Bacterial identification can be achieved either by examining protein mass spectral patterns or by using a subset of protein masses that are specific to the bacterium of interest. Several hundred protein components can be detected from a simple bacterial extract by off-line MALDI analysis of HPLC fractions. For *E. coli*, a few of them have been positively identified as *E. coli* proteins.[1] Yet, when comparing the mass table obtained by the MS method to the *E. coli* proteome database, it is found that only about half of the proteins in the mass table can be tentatively assigned to the known proteins based on their molecular masses. The proteins detected with masses matching with those in the proteome database are likely the products of gene expression from *E. coli*. The question remains: what are the origins of the proteins whose molecular masses do not match with those in the database? Identification of protein origins should be helpful in establishing the protein mass matching method as a scientifically valid method for bacterial identification, instead of just a number matching exercise.

There are several possible sources for the observed mass differences. The *E. coli* proteome database is quite extensive, compared to many other microorganisms. However, not all gene products have been identified.[2] Thus the proteins detected in the MS experiments may represent novel proteins that have not been found in the past and consequently are not listed in the database. The proteins detected in MS may also be modified proteins whose masses clearly are different from their unmodified counterparts. It is well known that multiple gene products (proteins) can be created from a single DNA sequence. The potential diversity of the final gene products is overwhelming considering that several hundred types of modifications that can alter proteins over time and in space (such as the locations within the cell). Although many different types of modification have already been identified, most of these modifications have not been localized on a sufficiently large number of proteins to allow the construction of extensive modification-specific databases that can reliably predict modifications from gene sequences. Therefore, analytical methods are required to identify not only the type of modification but also the site of modification in a protein. The observed protein modifications are listed in the proteome database, but it is by no means extensive. Finally, protein degradation *in vivo*, or fragmentation during sample preparation, may also account for the observed deviation of protein masses from those in the proteome database.

In this chapter, several proteins, including post-translationally modified proteins as well as *in vivo* protein fragmentation products, were identified by HPLC fractionation combined with multi-enzyme digestion and MALDI-TOF MS. During the course of this work, we unexpectedly discovered a possible mutated *E. coli* protein. Experimental results related to this effort are presented.

5.2 Experimental

Bacterial sample were cultured at the Edgewood RDE Center (ERDEC) using the same procedure as described in Section 2.2.1.[1, 3] Bacterial extracts were prepared by solvent suspension methods.[3] Chapter 3 has presented a more detailed discussion on this method. Proteins were separated by reverse phase HPLC using a preparative C_8 column (16×250 mm, Vydac, Hesperia, CA). The mobile phases were water (A) and acetonitrile (B) with 0.05% TFA in both phases. The gradient was 0-75% B over 300 minutes at a flow rate of 1 mL/min. Fractions were collected every minute during the run.

Trypsin, chymotrypsin and leucine aminopeptidase (LAP) were all purchased from Sigma (St. Louis, MO). About 0.05-0.1 μg trypsin or chymotrypsin was added to the 10 μL HPLC fractions, and adjusted to about pH 8 by 1 M NH_4HCO_3. Protein digestion was carried out at 37°C for 15 to 30 min. The exopeptidase LAP was used to digest the tryptic peptides as well as the intact proteins to obtain sequence information on the peptides and proteins. About 0.5-1 μg LAP was added to the tryptic digest or 10 μL fraction (pH 8). The N-terminal digests were sampled at 2 min, 5 min and 10 min by mixing the sample in 1:1 ratio (by volume) with the matrix solution. All N-terminal digestions were performed at room temperature.

MALDI mass spectra were collected by using a home-built time-lag focusing linear time-of-flight mass spectrometry as described in the previous chapters. Bovine ubiquitin, horse heart cytochrome c and their multiply charged species were used as internal standards for mass calibration. A two-layer method was used for MALDI sample preparation,[4] α-cyano-4-hydrooxycinnamic acid (HCCA) was used as matrix. In the two-layer method, the first layer was formed by applying 1 μL of 0.1 M HCCA in 20%

methenol/80% acetone to the MALDI probe tip and allowing it to dry very quickly in the air. For the second layer preparation, the sample solution was mixed 1:1 with saturated HCCA solution in 60% water/40% methanol. About 0.6-1 μL of the second-layer solution was then applied onto the first layer and allowed to dry. After this, 1 μL of distilled water was placed on the target for 10 s to wash away the salts before removing the water with a Kimwipe.

5.3 Results and Discussion

5.3.1 Identification of *E. coli* Proteins and Their Post-translational Modifications by Tryptic and Chymotryptic Peptide Mass Mapping

Trypsin is the most commonly used enzyme for peptide mass fingerprinting of proteins. It specifically cleaves the C-terminal side of lysine (K) and arginine (R) residues. A set of tryptic peptide masses from a sample are then matched against the theoretically generated tryptic peptide masses of proteins in a database. High-confidence protein identification can be routinely achieved from a purified protein or a simple protein mixture containing two or three proteins.[5-10] However, mass spectral peak intensities of the peptides from a tryptic digest are often found to be significantly different. One reason is that the absolute amount of each peptide generated from a protein is different due to the different cleave-off kinetics. In addition, some intrinsic properties of the peptides[11-16] as well as MALDI sample preparation methods[17] may account for the different peak intensities. It has been shown that a distinct suppression effect exists when MALDI is used to analyze tryptic peptide mixtures.[16] Figure 5.1 shows the effect of MALDI sample preparation on the tryptic digestion. Figure 5.1A was obtained by mixing the digest with the matrix solution in a 1:1 (by volume) ratio and using this mixture as the second layer in MALDI sample preparation. It is clear that except the peak at m/z 1232.7, all other peaks are very weak. No positive protein identification could be made based on the peptides observed in this spectrum. When the matrix to sample ratio was increased to 5:1 in the MALDI sample preparation, more peptide peaks were detected in the spectrum, as shown in Figure 5.1B. The protein was identified as 50S ribosomal protein L32 (RL32) using both the peptide mass mapping result obtained in Figure 5.1B and the measured molecular mass of the intact protein (i.e., 6315 Da). 93% of the protein sequence is covered by the peptide mass map, which provides a very high confidence for identification.

The protonated peptide peak at m/z 1232.7 corresponds to a peptide with the sequence of HHITADGYYR, while the second strongest peak at m/z 1041.4 in Figure 5.1B is from a peptide with the sequence of AVQQNKPTR. Although both peptides have an Arg (R) at the C-terminal end which could enhance the peptides signal intensity,[18] the former greatly suppressed the detection of the latter as shown in Figure 5.1A. This is likely due to their different hydrophobicities. The former is slightly more hydrophobic and therefore may more easily be incorporated into the hydrophobic HCCA matrix and thus is preferentially ionized. As the ratio between matrix and sample increases, the suppression effect is greatly reduced (Figure 5.1B). This is likely due to an increased probability for other peptides to form co-crystals with the more abundant matrix molecules, as well as an increase in the relative number of protons available from the matrix for ionization.

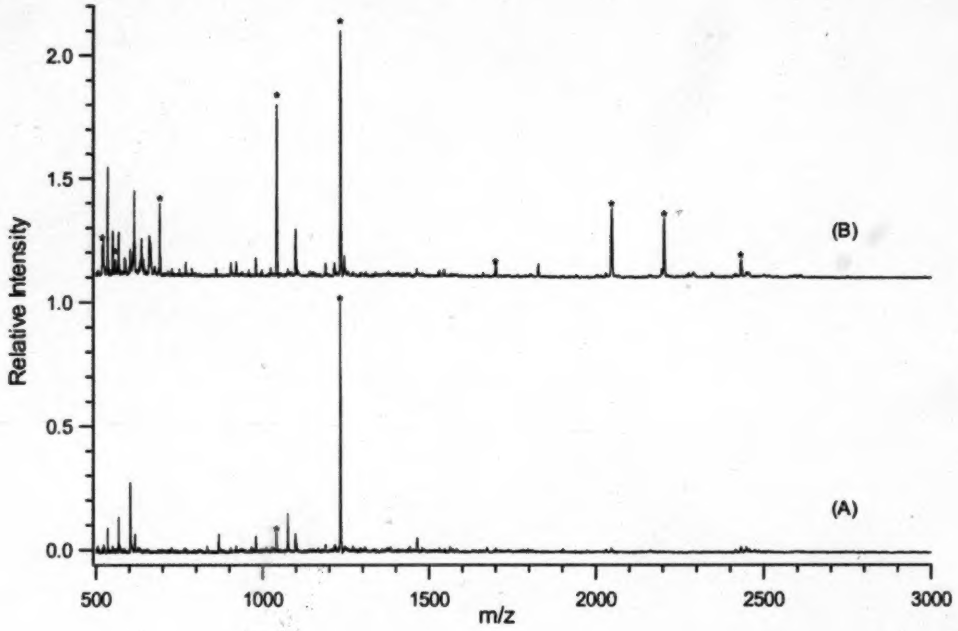

Figure 5.1 MALDI spectra of tryptic digestion on fraction #118 obtained by different second layer sample composition (A) tryptic digest mixed 1:1 (by volume) with the second layer matrix solution. (B) tryptic digest mixed 1:5 with the second layer matrix solution.

As mentioned earlier, poor sequence coverage by MALDI analysis of tryptic digest could also be the consequence of enzyme cleavage kinetics. That is the concentrations of some tryptic fragments in the digest solution are too low to be detected by MALDI. In these cases, optimization of sample preparation method usually cannot significantly increase the sequence coverage. An alternative approach is to use other enzymes with different specificities. In this work, we investigated the applicability of using chymotrypsin for peptide mass mapping. Chymotrypsin cleaves the C-terminal side of some hydrophobic residues. As shown below, it can provide complementary information to tryptic digestion and the combination of the two digestions can be very useful for unambiguous protein identification.

Figure 5.2 shows the MALDI spectrum of fraction #114 and its peptide mass mapping spectra from tryptic and chymotryptic digestions. The MALDI spectrum shown in Figure 5.2A indicates that there is one major component in this fraction with MH^+ at 5381 Da and the doubly charged species of this component is also detected in the MALDI spectrum. The tryptic peptide masses were used to search for the protein(s) in the database using MS-Fit in the UCSF Protein Prospector searching program. *E. coli* 50S ribosomal protein L34 came out as the top candidate with an amino acid sequence coverage of 76%.

It should be noted that an individual fraction from the HPLC separation of the cell extract often contains a mixture of several proteins. The relative signal intensities in a MALDI spectrum do not reflect the relative amounts of proteins in a mixture. Thus, the major protein component in the fraction may give a weak signal in the MALDI spectrum. Therefore, the tryptic peptides generated from the digestion of an individual HPLC fraction may be from the protein that shows a weak signal in the MALDI spectrum of the fraction. For example, for fraction #114, the tryptic peptides at m/z of 1232 and 1257 might be from the protein with MH^+ at 6316 Da, as calculated from theoretical digestion of 50S ribosomal protein L32 (MH^+ 6316 Da). This protein generates a peak with relatively low intensity in the MALDI spectrum of fraction #114, whereas it gives a dominant peak in fraction #118. Figure 5.2C shows the peptide mass map from

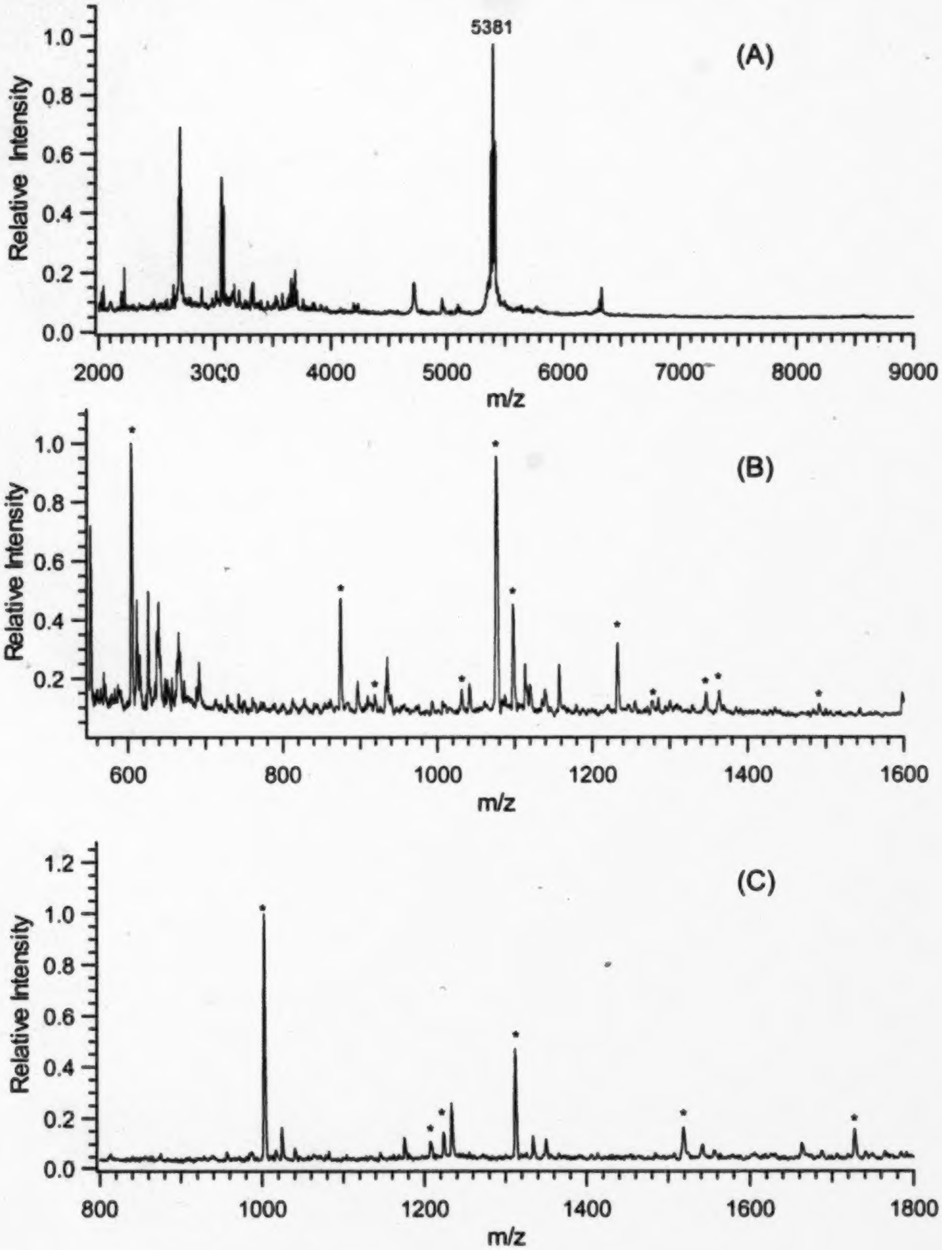

Figure 5.2 MALDI spectra of fraction #114 and its tryptic and chymotryptic digests. (A) molecular weight determination, (B) tryptic digestion, (C) chymotryptic digestion. Peaks marked with * represent those peptides matched with the theoretical digestion.

chymotrypsin digestion on fraction #114. Six peaks matched the theoretical chymotryptic digest of 50S ribosomal protein L34. 100% protein sequence was covered by the chymotryptic mass map. Combining the results from the two enzymatic digestion experiments, we can conclude with confidence that the major component in this fraction is 50S ribosomal protein L34.

Protein modifications, occurring either during the cell cycle (co- and post-translationally modified) or during sample treatments (i.e., artificially modified), will result in a deviation of the detected protein mass from that in the proteome database that is mostly predicted on the basis of the DNA sequence. To identify such modifications and to localize the modification sites, a high protein sequence coverage from the peptide map is very important. This may be achieved using multiple digestions with enzymes of different specificities. Figure 5.3 shows the MALDI spectrum of fraction #124 displaying a major peak with MH^+ at 6255 Da. Table 5.1 lists the theoretical and observed tryptic peptides. Peptide mass mapping gives a match with *E. coli* 50S ribosomal protein L33 (83% coverage). However, the predicted molecular mass of this protein is 6240 Da, which is 14 Da lower than the observed one, suggesting a modification. Most probably methylation has occurred to this protein by either an *in vivo* or *in vitro* process.

Table 5.2 shows the result from the chymotrypsin digestion. The observed MH^+ for chymotryptic peptides residue 1-10 and residue 1-20 are 14 Da higher than the theoretically calculated ones, indicating that the modification occurred somewhere between residue 1-10. We note that it has been reported that several *E. coli* ribosomal proteins have N-methylated amino-terminal residues; for example, ribosomal protein L11 has N-Me_3Ala,[19, 20] L16 has N-Me-Met,[21] and L33 has N-Me-Ala.[22]

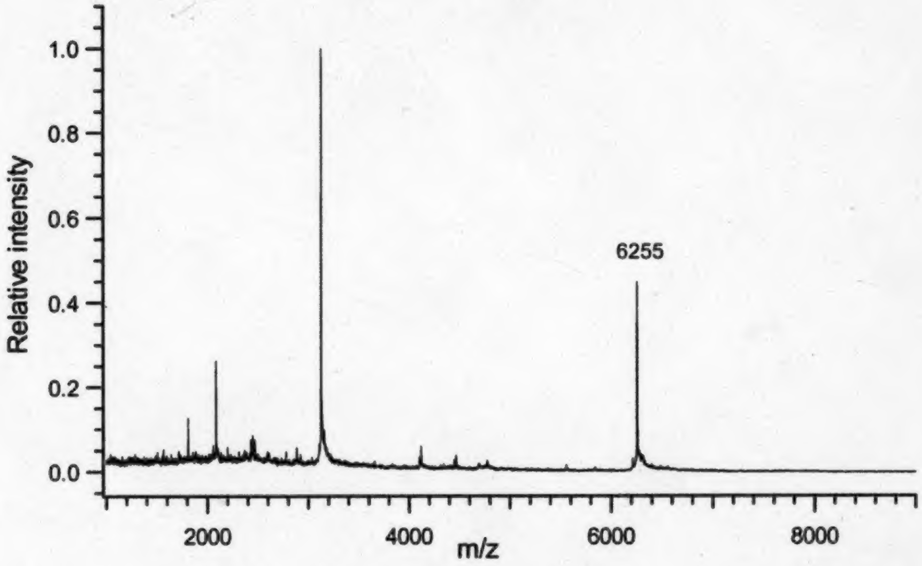

Figure 5.3 MALDI spectrum of fraction #124 shown a major component of MH^+ at 6255 Da.

Table 5.1 Tryptic peptides matched 50S ribosomal protein L33 from fraction #124.

	Protonated Mass $[M+H]^+$	
Tryptic Fragments	Theoretical Fragments	Observed Fragments
44-49	787.4	787.4
37-43	860.5	860.7
28-36	1085.7	1086.1
44-52	1115.6	1116.1
27-36	1241.8	1242.1
10-24	1569.7	1569.5
37-49	1628.9 *	1628.8*
8-24	1811.0 *	1810.3*
10-27	1968.0*	1968.6*

*Average protonated mass

Table 5.2 Chymotryptic peptides matched 50S ribosomal protein L33 from fraction #124.

	Protonated Mass $[M+H]^+$	
Chymotryptic Fragments	Theoretical Fragments	Observed Fragments
1-10	**1155.6**	**1170.0**
39-48	1225.7	1225.7
21-33	1545.8	1546.2
36-48	1630.0	1630.1
34-48	1872.2*	1872.2*
1-20	**2163.5***	**2177.5***

*Average protonated mass

Table 5.3 Summery of the identification of proteins from HPLC fractions by dual enzyme digestion.

Fraction #	Observed (MW, Da)	Trypsin Digest (% sequence coverage)	Chymotrypsin Digest (%sequence coverage)	Identity (MW, Da)
108	5111.5	7 peaks match (80%)	6 peaks match (100%)	30S Ribosomal Protein S22 (Met-ox)(5111)
114	5380.6	11 peaks match (76%)	6 peaks match (100%)	50S Ribosomal Protein L34 (5380)
118	6315.0	11 peaks match (93%)	10 peaks match (100%)	50S Ribosomal Protein L32 (6315)
124	6254.1	10 peaks match (83%)	8 peaks match (100%)	N-terminal Methylated 50S Ribosomal Protein L33 (6254)
151	7706.6	16 peaks match (99%)		Chain Hypothetical Protein YAHO at position 22-91 (7707)
208	9226.0	11 peaks match (76%)	4 peaks match (92%)	DNA-binding Protein HU BETA (NS1) (HU-1) (9226)

Using the dual enzyme digestions, several other proteins from the HPLC fractionation of *E. coli* 9637 were positively identified. The results are summarized in Table 5.3. In addition to the methylation of 50S ribosomal protein L33, we detected one of the most frequently occurred modifications, the methionine oxidation. The oxidized forms are observed for all the proteins listed in Table 5.3 along with their non-oxidized forms. In some cases, the oxidized species can give the dominant peaks. For example, for 30S ribosomal protein S22 shown in Table 5.3, only a very weak peak at $MH^+ = 5096$ Da was detected. The sulfur atom in methionine is chemically very reactive, particularly toward oxdizing agents. Oxidation to form sulfoxide is commonly observed in protein samples stored *in vitro* in the absence of antioxidants, and in certain cases, it is also observed *in vivo*.[23]

The protein with MH^+ at 7708 Da was identified as Chain Hypothetical Protein YAHO with residue 22-91, which is a part of the larger sequence for Hypothetical 9.9 kDa protein in bett-prpr intergenic region (P75694). The cleavage of a short piece of N-terminal signal peptide (residue 1-21) most likely occured when the protein was exported through the inner membrane.[24]

5.3.2 N-terminal Digestion Using LAP

In this work, exopeptidase LAP was used to generate a stretch of amino acid sequence from the intact proteins. The stretch of amino acid sequence has been termed as "protein terminal sequence tag" and has been proposed as a valuable attribute for protein identification.[25, 26] It is interesting to note that for microorganisms with small genomes, the protein terminal sequence tag can be very specific. For example, about 60% proteins in *E. coli* have unique N-terminal sequence tags of length 4 amino acids.[24] For proteins that do not have unique N-terminal sequence tag, there are relatively few proteins which share the same tag. The most frequent N-terminal tag of length 4 amino acids

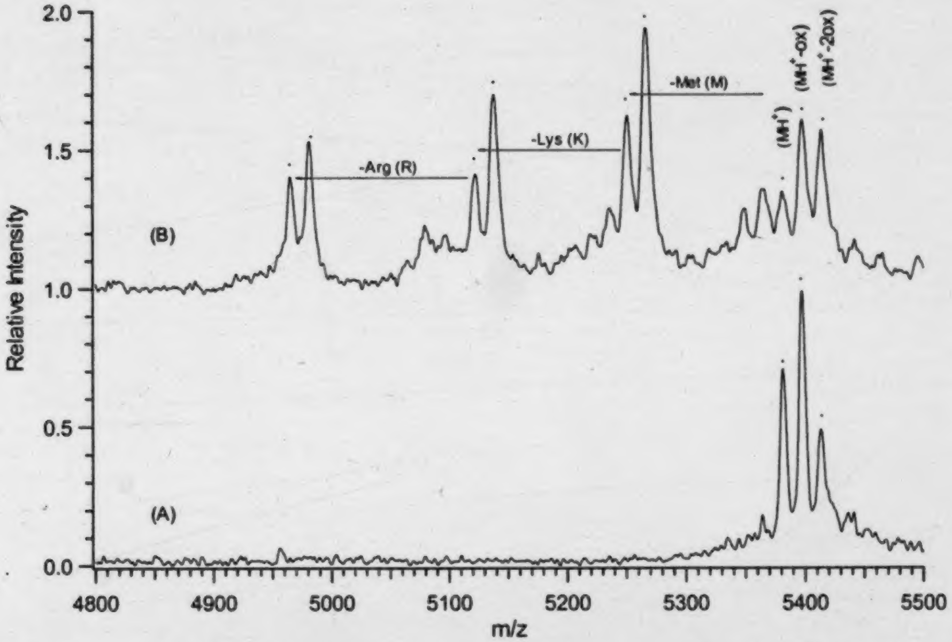

Figure 5.4 LAP N-terminal digestion of intact protein from fraction #114. (A) MALDI spectrum showing the molecular ion and its methionine oxidized ions, (B) MALDI spectrum taken after 2 min LAP N-terminal digestion.

(MKTL) was found to be only 10 proteins.[27] Most protein terminal sequence tag to date has been generated by the N-terminal Edman degradation method, which has been the method of choice for protein identification until recently. Edman degradation is a time-consuming technique that requires a relatively large amount of material. By contrast, combining exopeptidase digestion and MALDI analysis, a short sequence tag can be obtained much more rapidly and sensitively.[25]

Figure 5.4 is the result from N-terminal digest of fraction #114 by LAP. After 2 min, three new peaks appeared. As the digestion goes further, the relative intensity of the original molecular ion peak gets weaker, while the new peaks become stronger. The short stretch of amino acids MKR matches with the N-terminal amino acid sequence of 50S Ribosomal Protein L34.

During the course of our work using N terminal digestion combined with MALDI for protein identification, we found a conflicting result for fraction #118 through the N terminal digestion of the intact protein. Peptide mass mapping results from both trypsin and chymotrypsin digestion have identified the protein to be 50S ribosomal protein L32. However, the LAP digestion of Fraction #118 gives a N-terminal tag of QA, as shown in Figure 5.5, instead of AVQ as predicted from the *E. coli* genome. It seems there are some amino acid position switches among the amino acid residues at the N-terminal side, though the switches were not manifested by either trypsin or chymotrypsin digestion. N-terminal digestion of the tryptic peptides was further performed in order to confirm that the protein in this fraction is indeed a mutated form of 50S ribosomal protein L32. The experimental result is shown in Figure 5.6. After 2 min LAP digestion, 3 new peaks were detected, corresponding a peptide sequence tag of HHIT. This is consistent with the expected partial sequence of peptide at MH^+ 1232.3 Da. The N-terminal sequence of the peptide with MH^+ at 1041 Da could not be obtained by LAP digestion. This is not surprising, as having discussed before, the peptide with MH^+ at 1232 Da is always preferentially detected from the tryptic peptide mixture by MALDI. It should be noted that the biological consequence, if any, of the mutation found for the protein 50S ribosomal protein L32 in this particular sample remains to be determined.

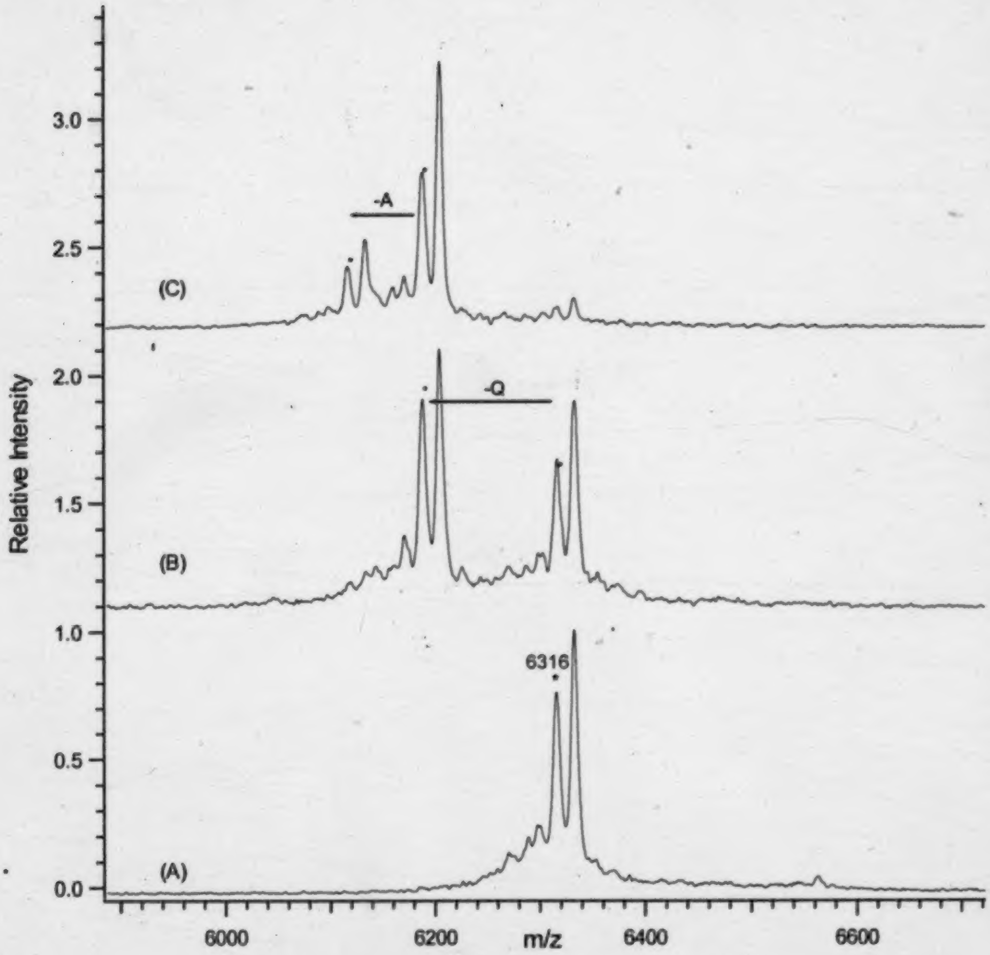

Figure 5.5 LAP N-terminal digestion on intact protein in fraction #118. (A) MALDI spectrum of fraction # 118, (B) The spectrum taken after 2 min LAP digestion, (C) The spectrum taken after 5 min LAP digestion.

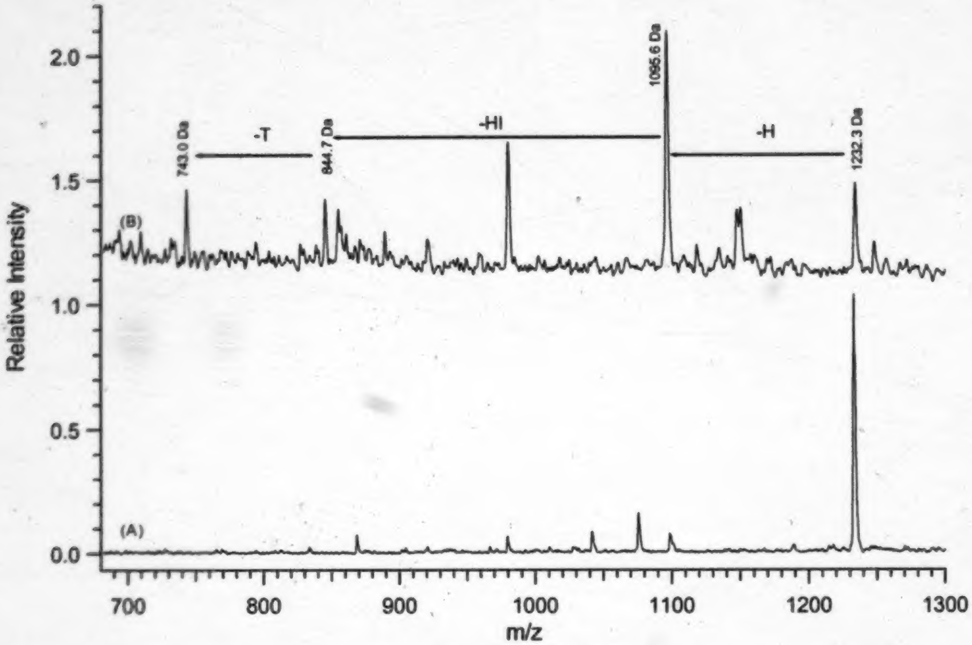

Figure 5.6 LAP N-terminal digestion on tryptic digest of fraction #118. (A) Tryptic digestion result showing the major peak with m/z at 1232.3; (B)The spectrum taken after 5 min LAP digestion on the tryptic digest of fraction #118.

5.4 Conclusions

In this chapter, several proteins, including some post-translational modified proteins and protein degradation products, were successfully identified from an *E. coli* extract using HPLC fractionation, followed by multiple-enzyme digestion and MALDI TOF MS analysis. These results provide direct evidence that the proteins detected by MALDI include those with modifications. The mass spectrometric approach presented herein should be very useful to provide information on protein modifications that can be incorporated into the database, thereby enhancing the utility of the current proteome database for bacterial identification as well as biological applications (e.g., functional studies). From the protein identities, it is clear that those commonly observed peaks in

MALDI spectra are an subset of proteins that generally exist in high abundance inside the cells. Moreover, these proteins are generally very basic and thus are easily ionized in the positive ion MALDI analysis. Since several hundred proteins exist in the cell extract, it is impossible to separate them into pure HPLC fractions using an one-dimensional HPLC separation. For most HPLC fractions, more than five proteins co-exist, and in most cases, these proteins are in dramatically different concentrations. Therefore, it is not possible to identify them by peptide mass mapping alone. To achieve extensive protein identification, MS/MS is required such that proteins can be identified by using peptide sequence information. The use of LC MS/MS of tryptic peptides in combination with HPLC fractionation of bacterial extract will be discussed in Chapter 7.

5.5 Literatures Cited

1. Dai, Y.; Li, L.; Roser, D. C.; Long, S. R. *Rapid Commun. Mass Spectrom.* **1999**, 13, 73.
2. Blattner, F. R.; Plunkett, G. III.; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A. Goeden, M. A.; Rose. D. J. Mau, B.; Shao, Y. *Science*, **1997**, 277, 1453.
3. Wang, Z.; Russon, L.; Li, L.; Roser, D. C.; Long, S. R. *Rapid Commun. Mass Spectrom.* **1998**, 12, 456.
4. Dai, Y.; Whittal, R. M.; Li, L. *Anal. Chem.* **1996**, 68, 2494.
5. Cleveland, D. W.; Fischer, S. G.; Kirschner, M. W.; Laemmli, U. K. *J. Biol. Chem.* **1977**, 252, 1102.

6. Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimlay, C.; Watanabe, C. *Proc. Natl. Acad. Sci.* USA, **1993**, 90, 5011.

7. James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. *Biochem. Biophys. Res. Commun.* **1993**, 195, 58.

8. Mann, M.; Hojrup, P.; Roepstorff, P. *Bio. Mass Spectrom.* **1993**, 22, 338.

9. Pappin, D. J. C.; Hojrup, P.; Bleasby, A. *J. Curr. Biol.* **1993**, 3, 327.

10. Yate, J. R. III.; Speicher, S.; Griffin, P. R.; Hunkapiller, T. *Anal. Biochem.* **1993**, 214, 397.

11. Cohen, S. L.; Chait, B. T. *Anal. Chem.* **1996**, 68, 31.

12. Zhu, Y. F.; Lee, K. L.; Tang, K.; Allman, S. L.; Taranenko, N. I.; Chen, C. H.; *Rapid Commun. Mass Spectrom.* **1995**, 9, 1315.

13. Valero, M. L.; Giralt, E.; Andreu, D.; In *Peptides* 1996; Ramage, R.; Epton, R.; Eds.; Mayflower Scientific Ltd.; Kingswinford, UK, **1998**; pp855-856.

14. Olumee, Z.; Sadeghi, M.; Tang, X. D.; Vertes, A. *Rapid Commun. Mass Spectrom.* **1995**, 9, 744.

15. Wenschuh, H.; Halada, P.; Lamer, P.; Jungblut, P.; Krause, E.; *Rapid Commun. Mass Spectrom.* **1998**, 12, 115.

16. Kratzer, R.; Eckerskorn, C.; Karas, M.; Lottspeich, F. *Electrophoresis* **1998**, 19, 1910.

17. Kussmann, M.; Nordhoff, E.; Rahbek-Nielsen, H.; Haebel, S.; Rossel-Larsen, M.; Jakobsen, L.; Gobom, J.; Mirgodskaya, E.; Kroll-Kristensen, A.; Palm, L.; Roepstorff, P. *J. Mass Spectrom.* **1997**, 32, 593.

18. Krause, E.; Wenschuh, H.; Jungblut, P. R. *Anal. Chem.* **1999**, 71, 4160.

19. Lederer, F.; Alix, J. H.; Hayes, D. *Biochem. Biophys. Res. Commun.* **1977**, 77, 470.

20. Dognin, M. J.; Wittmann-Liebold, B. *Eur. J. Biochem.* **1980**, 112, 131.

21. Brosius, J.; Chen, R. *FEBS Lett.* **1976**, 68, 105.

22. Wittmann-Liebold, B.; Pannenbecker, R. *FEBS Lett.* **1976**, 68, 115.

23. Allen, G. *Protein*, JAI Press, London, England, **1997**, Volume 1, 12.

24. Nielson, H.; Engelbrecht, J.; Brunak, S.; Von Heijne, G. *Protein Eng.* **1997**, 10, 1.

25. Wilkins, M. R.; Gasteiger, E.; Sanchez, J. C.; Appel, R. D.; Hochstrasser, D. F. *Curr. Biol*, **1996**, 6, 1543.

26. Gooley, A. A.; Ou, K.; Russell, J.; Wilkins, M. R.; Sanchez, J. C.; Hochstrasser, D. F.; Williams, K. L. *Electrophoresis*, **1997**, 18.

27. Willkins, M. R.; Willams, K. L.; Appel, R. D; Hochstrasser, D. F. (Eds.) Springer, In *Proteome Research: New Frontiers in Functional Genomics*, **1997**, 41.

Chapter 6

Nanoliter Protein Concentration and Digestion Combined with Microspot MALDI MS for Identification of Proteins Fractionated by Conventional HPLC[a]

6.1 Introduction

In Chapter 5, we identified several bacterial proteins and their post-translational modifications by HPLC fractionation of a bacterial extract, followed by multi-enzyme digestion and MALDI analysis. A preparative HPLC column was used to fractionate the crude cell extract in order to isolate a sufficient amount of protein for enzyme digestion and subsequent analysis of the digest by MS. Unfortunately, preparative HPLC usually has poor resolving power and it requires a large amount of starting material. The later can be a major limitation for proteomics projects involving a limited supply of cells such as those from tumor tissue of a patient. Analytical or small-bore column HPLC provides much better resolution and requires a significantly smaller amount of starting material. However, the fractions collected by analytical column separation were often too dilute to allow peptide mass fingerprinting by enzymatic digestion. To circumvent this problem, one can carry out multiple runs and then pool the corresponding fractions together for subsequent protein identification. However, multiple HPLC runs are naturally very time consuming.

Recently, Lubman and coworkers[1] have reported an elegant technique that combines the benefit of high sample loading of preparative HPLC with the high resolving power of analytical HPLC for MS separation and identification of proteins. High sample loading and high resolving power are achieved by using two columns with different lengths linked in series, and by holding them at different temperatures during separation. The column length and temperature of the first column are optimized to provide high sample loading with some resolution for the protein mixture. The conditions for the second column are optimized to provide high resolution for the proteins. While this technique provides an alternative to running multiple analytical-column HPLC, the amount of starting material required for protein identification is still equivalent to that used in a preparative HPLC experiment.

---

[a] A form of this Chapter is in preparation for publication: B. O. Keller, Z. Wang, L. Li *"Bacterial Protein Identification by HPLC Fractionation, Nanoliter Digestion and Microspot MALDI Analysis"*. Dr. B. O. Keller collected the mass spectra of the standard proteins.

In this chapter, a method that allows multiple experiments to be carried out from individual fractions separated using analytical-column HPLC is presented. This is made possible by using a nanoliter sample handling technique, as opposed to conventional microliter volume experiments. In this work, we describe a technique for pre-concentrating a protein solution inside a capillary tube, followed by chemical and enzymatic reactions. The resulting peptides are analyzed by microspot MALDI. The performance of the technique is demonstrated in the characterization of protein fractions originating from analytical HPLC column fractionation of *E. coli* extracts.

6.2 Experimental

6.2.1 Chemicals and Materials

E. coli bacteria samples were from Edgewood RDE Center, Aberdeen Proving Ground, MD, USA. Dithiothreitol (DTT), iodoacetamide, α-cyano-4-hydroxycinnamic acid (HCCA) and trypsin (98%, L-1-Tosylamide-2-phenylethyl chloromethyl ketone

(TPCK) treated for reduction of chymotrypsin activity), horse cytochrome c, leucine aminopeptidase (LAP), and trifluoroacetic acid (TFA) were from Sigma-Aldrich-Fluka (Oakville, Ontario, Canada). HCCA was recrystallized from ethanol (95%) at 50C before use.

6.2.2 Extraction of Bacterial Proteins

The *E. coli* 9637 extract was prepared by solvent suspension methods, which is described in detail in Chapter 2.

6.2.3 HPLC Fractionation

Separation of *E. coli* 9637 extract was performed on a HP1100 HPLC (Hewlett-Packard, Palo Alto, USA) using a 4.6×250 mm C_8 column (Vydac, Hesperia, CA). The mobile phases were nanopure water (A) and acetonitrile (B) with 0.05% TFA in both phases. The solvent gradient was 2-20% B over 10 min, 20-40% B over 40 min, and then 40-55% B over 10 min. The flow rate was 0.5 mL/min. Fractions were collected every minute during the run. For each separation, 30 μL of the *E. coli* extract containing 5 mg starting lyophilized *E. coli* sample was injected into the C_8 column. The individual fraction was concentrated to about 10 μL by a high-speed vacuum centrifuge.

6.2.4 In-capillary Sample Concentration, Reaction and Microspot MALDI Sample Preparation

The nanoliter chemistry station has been described in more detail elsewhere.[2] The polyimide coating was burned off from one end of a short piece of fused silica capillary (20 μm I.D, 10-15 cm in length) (Polymicro Technologies, Phoenix, Arizona, USA). The end was etched by 45% HF with nitrogen continuously flowing through the capillary tube. The capillary was then connected to a syringe and used to draw sub-nanoliter volumes of protein sample from a horizontally mounted pipette tip. To minimize analyte loss due to irreversible adsorption onto the wall surface, the capillary was treated with a siliconizing agent before use (Glassclad-18, United Chemical Technologies, Bristol, PA, USA). The sample plug was observed under a microscope with 40 or 60× magnification, its volume was determined using a calibrated recticle that was positioned in the eyepiece. For in-capillary sample concentration, a ~500 pL sample plug was dried inside the capillary close to the capillary entrance (Figure 6.1). This step can be repeated many times to achieve sufficient sample concentration inside the capillary. To accelerate the drying process, an orthogonal N_2 gas was applied at the open capillary end. After 1 or 2 concentration steps, a plug of ~1 nL of triply

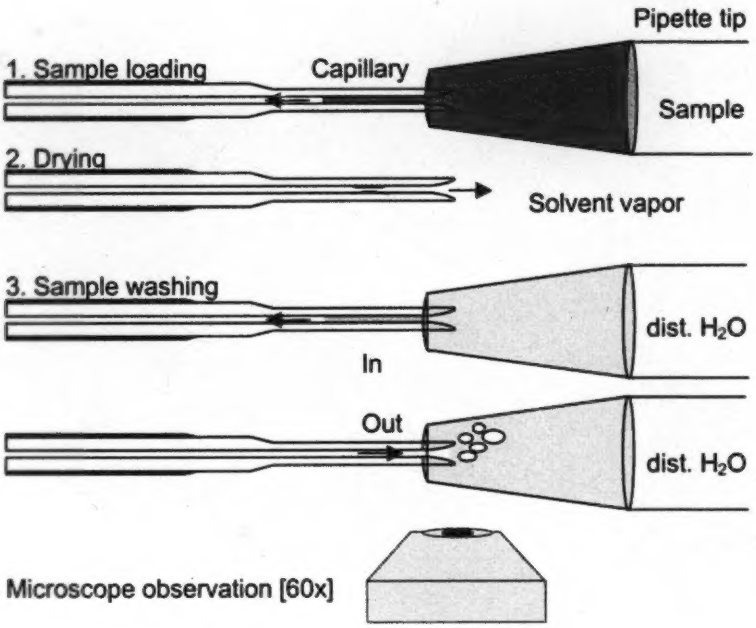
Pipette tip
1. Sample loading
Capillary
Sample
2. Drying
Solvent vapor
3. Sample washing
dist. H₂O
In
Out
dist. H₂O
Microscope observation [60x]

Figure 6.1 Schematic drawing of in-capillary sample concentration and washing steps.

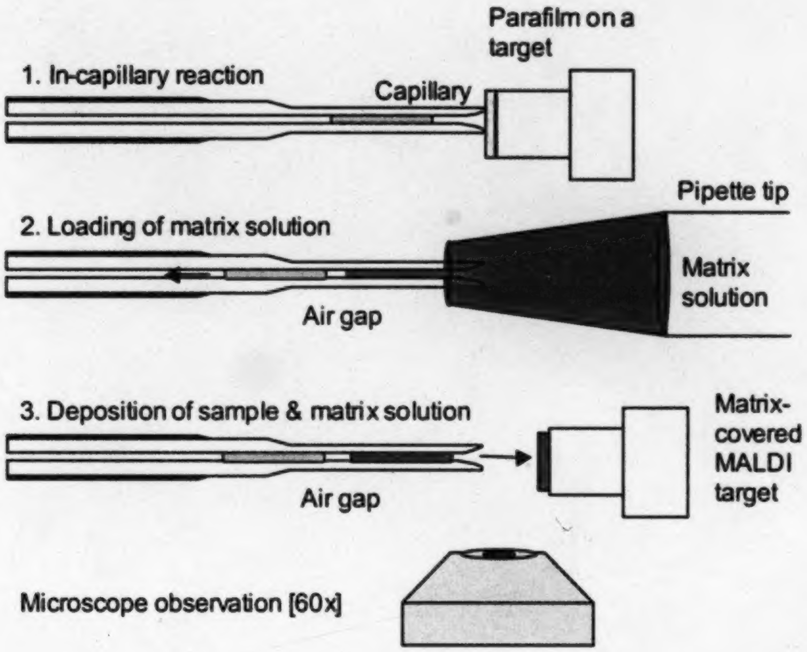
Parafilm on a target
1. In-capillary reaction
Capillary
Pipette tip
2. Loading of matrix solution
Matrix solution
Air gap
3. Deposition of sample & matrix solution
Matrix-covered MALDI target
Air gap
Microscope observation [60x]

Figure 6.2 Schematic drawing of in-capillary reaction and microspot sample deposition.

distilled water was drawn into the capillary and pushed out after ~20 sec. This is critical since the small capillary is easily plugged by the accumulation of excessive salts. After the final washing step, enzyme or other chemical solutions were drawn into the capillary (Figure 6.2). The capillary was then pushed against a piece of Parafilm to close the entrance and thus stop any further evaporation. After sufficient reaction time the sample/enzyme or sample/chemical mixture was again dried inside the capillary and further chemical or enzymatic reactions were performed by introducing different chemical/enzyme solutions in an additional step. When all desired reaction steps had been performed a ~500 pL plug of saturated matrix solution was drawn into the capillary. The sample and matrix solution were separated by a small air gap. Both plugs were then simultaneously deposited from an approximate 0.1 mm distance onto a matrix-covered MALDI target.

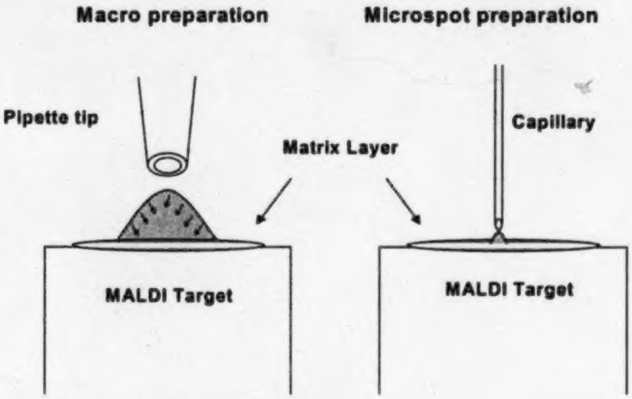

Figure 6.3 Schematic comparison of macro- and micro-MALDI sample preparation.

To prepare the matrix-covered MALDI target, about 1 µL of a 5 mg/mL solution of HCCA in 80% acetone/methanol (v/v) was first deposited on the clean probe to form the thin first layer. A second layer of 0.4 µL of HCCA saturated in 35% methanol/water (v/v) was deposited onto the first layer and allowed to dry. Compared to the commonly used MALDI sample preparation methods, the advantage of microspot sample preparation is obvious. As illustrated in Figure 6.3, commonly used microliter sample preparation results in sample spots of several mm in diameter, while microspot sample deposition by a small capillary yields sample spots of 80 to 200 µm in diameter. The dramatic reduction of spot size acts as an *in situ* sample concentration step on the MALDI target, thereby greatly increasing the analysis sensitivity.[3]

6.2.5 MALDI Analysis

Mass spectra of proteins and their digests were collected on a home-built linear time-lag focusing MALDI-TOF mass spectrometer, equipped with a 337 nm laser having a 3 ns pulse width (model VSL 337ND, Laser Sciences Inc., Newton, MA, USA). This home-built instrument has been described in detail elsewhere.[4] In general, 150-200 laser shots (3-5 μJ pulse energy) were averaged to produce a mass spectrum. Spectra were acquired and processed with Hewlett-Packard supporting software and reprocessed with the Igor Pro software package (Wavemetrics, Inc., Lake Oswego, OR). Each spectrum was normalized using the most intense signal.

6.3 Results and Discussion

6.3.1 Effects of In-capillary Concentration and Cleaning Steps

When dealing with diluted protein samples, such as those fractions from an analytical-column separation, an additional concentrating step is usually necessary to get sufficient protein concentration for efficient enzymatic digestion. The concentration can be done inside the capillary as shown in Figure 6.1. However, capillary blockage due to the extensive accumulation of salt contaminants became a major problem during the in-capillary concentration step, since the capillary used here has a very small internal diameter (20 μm). This problem can be solved by washing the dried protein sample using distilled water after every one or two concentration steps. It is worthy to mention that the simultaneous deposition of protein digest with matrix solution serves as an *in situ* sample cleaning step for MALDI analysis. These small scale sample cleaning procedures are very important since it is not practical to perform any on-probe washing step for

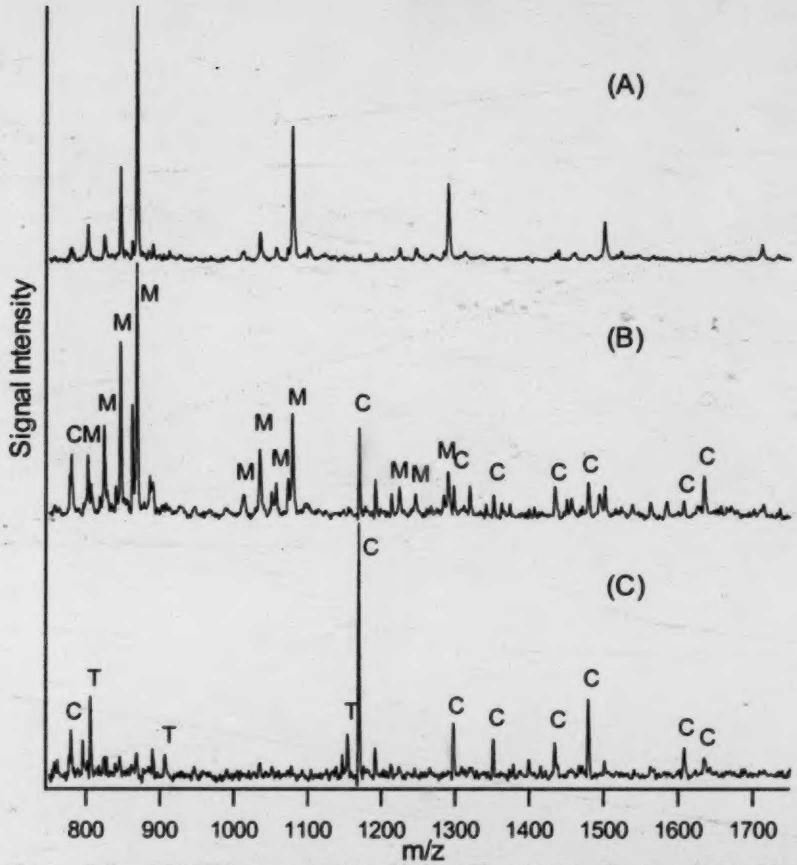

Figure 6.4 MALDI mass spectra of in-capillary tryptic digests of a 4 μM cytochrome C solution in 40 mM NaCl and 20 mM NH_4HCO_3 buffer. (A) Direct deposition of digest mixture onto MALDI target without any washing step. (B) Simultaneous deposition of contaminated digest mixture and a matrix solution plug onto target. (C) Simultaneous deposition of digest mixture of washed protein and matrix solution onto target. Peaks marked with C are tryptic peptides from cytochrome c; peaks marked with M are matrix clusters; peaks marked with T are trypsin autolysis peptides.

microspot MALDI sample preparation, this is because the tiny amount of deposited protein sample would be easily washed off.

The effects of these two cleaning steps are demonstrated in Figure 6.4. A highly salt-contaminated cytochrome c solution was dried inside the capillary and digested by trypsin. The spectrum of the untreated digest mixture (Figure 6.4A) shows mainly matrix-cluster peaks, which is a common observation for protein samples with such a high salt content.[5] Note that the digestion was done at a relatively low concentration of digestion buffer (only 20 mM NH_4HCO_3). Generally, higher concentrations of up to 100 mM NH_4HCO_3 are used because optimal activity of the enzyme is at a higher pH. However, direct deposition of protein digest in such a high buffer concentration would easily dissolve the pre-deposited matrix layer. The second spectrum (Figure 6.4B) shows matrix, clusters with eight tryptic peptides of cytochrome c discernable from the matrix clusters. The dramatic improvement is because tryptic peptides co-crystallize with matrix easily while the less hydrophobic components such as salts tend to be pushed into the rim of the sample spot and are excluded from the crystal. Another practical aspect of simultaneous deposition of sample with matrix solution is that it neutralizes the basic buffer solution before deposition onto the matrix covered MALDI target. This is very critical since the direct deposition of digest with a high basic buffer content is usually not feasible. The matrix layer on the target would be immediately dissolved and thus the experiment ruined. The best spectrum was obtained by the combination of the two cleaning steps. This is shown in Figure 6.4C. A very clean MALDI spectrum with no matrix cluster interferences was observed.

Using the concentration and washing setup shown in Figures 6.1, proteins are retained onto the capillary by hydrophobic or other nonspecific interactions with the C_{18} coating. Contaminants, such as salts, do not strongly interact with the capillary wall. Therefore they are easily washed out with the distilled water. This concentration and washing step is somewhat comparable to other, larger scale protein cleanup procedures such as C_{18}-coated microbeads in pipette tips (e.g., ZipTips).[6,7] The difference in our case is that the protein sample was dried inside the capillary. This ensures a more complete transfer of the proteins to the capillary wall and potential wall binding sites.

6.3.2 Bacterial Protein Identification by In–capillary Nanoliter Digestion

The nanoliter sample handling technique described above was applied to identify bacterial proteins fractionated by conventional HPLC. It should be pointed out that the fractions were pre-concentrated by SpeedVac from the original volume of 500 μL to about 10 μL shortly after the fractionation experiment. Fractions need to be stored in the freezer. Also it was found that concentrated fractions would suffer less from loss to the container wall during long-term storage. After pre-concentration, the protein concentrations in some of the fractions might be sufficiently high for conventional in-solution digestion. However, using conventional microliter sample preparation technique, only a few experiments can be done with 10 μL of sample. In contrast, with the nanoliter sample preparation method, only a few nanoliters of sample is used for each experiment. Thus, many experiments, including optimization of digestion conditions and digestion with different enzymes, can performed from a microliter fraction. In our experiments, usually a few nanoliters of sample was first taken for molecular weight analysis, followed by trypsin digestion for peptide mapping. If peptide mapping along with the molecular weight information cannot unambiguously identify the protein in the database, several nanoliters from the remaining fraction were taken for further experiments until we could confidently identify the protein or determine that identification is not possible with the currently available techniques and database.

Figure 6.5 shows one example where *E. coli* 30S ribosomal protein S20 (P02378, MW 9554 Da) was positively identified by peptide mass mapping in combination with

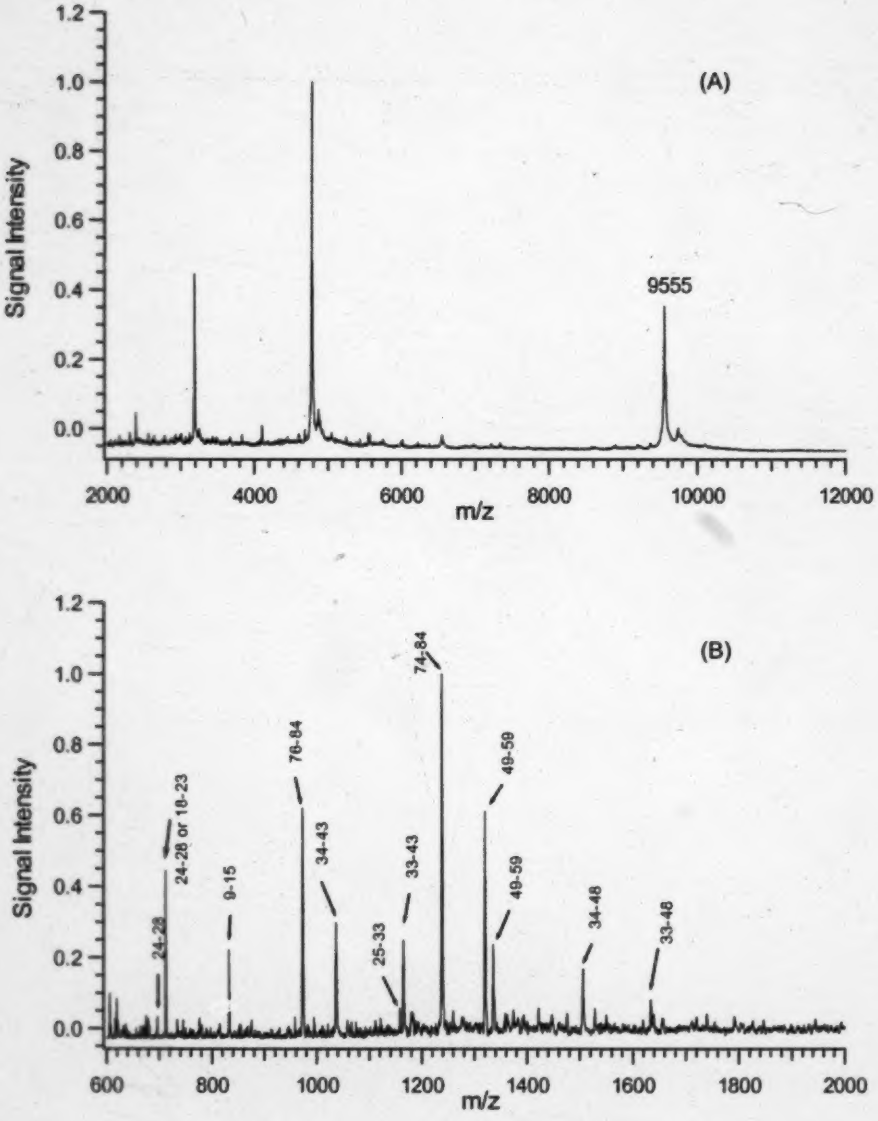

Figure 6.5 MALDI analysis of HPLC fraction #42 of *E. coli* extract. (A) Molecular weight determination, (B) MALDI peptide mass mapping. The amino acid sequences covered are labeled.

accurate molecular weight determination. A very clean peptide mass map was obtained. Almost all the major peaks matched 30S ribosomal protein S20 with a sequence coverage of 60%. Note that the molecular weight of this protein is 9684 Da in the proteome database, the one detected in Figure 6.5A has lost its N-terminal methionine.

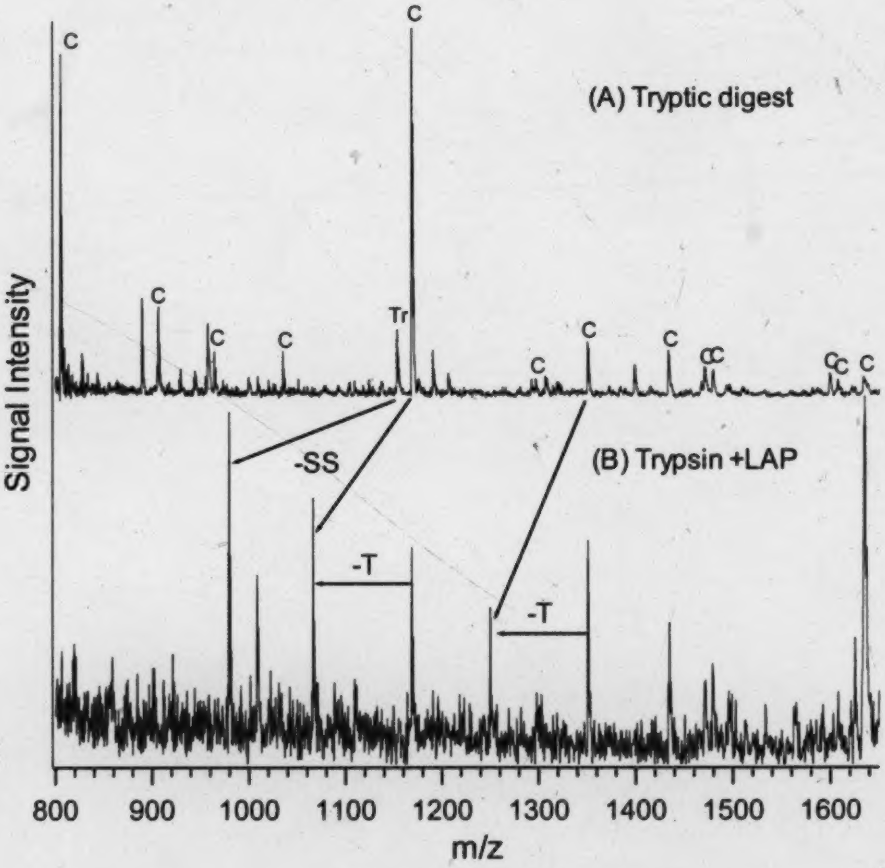

Figure 6.6 MALDI mass spectra of in-capillary digests of cytochrome c.

Although peptide mass mapping of the protein shown in Figure 6.5 was adequate for identification of the protein in question, many cases exist where an insufficient number of peptides were detected for an unambiguous protein identification by peptide mass mapping alone. This could either be due to a lack of sufficient starting material to generate an adequate number of peptides, or due to contamination with other peptides, which obscure the database search results. It has been shown by different groups that additional sequence information of only one or two tryptic peptides (i.e., sequence tag) is often enough for confident protein identification.[8,9] A common technique for obtaining sequence information is to use a tandem mass spectrometer to generate MS/MS fragment ion spectra of peptides. MS/MS spectra can be obtained by using collision-induced dissociation (CID) in tandem mass spectrometry[10,11]. Alternatively, post-source decay (PSD) fragment ion spectra can be obtained using a reflectron MALDI TOF instrument.[12]

A different way of obtaining additional sequence information is the application of exoproteolytic enzymes directly on tryptic digest mixtures. Our lab has used trypsin digestion followed by aminopeptidase or carboxyopeptidase to obtain sequence information on tryptic peptides. This method has been used to locate the modification site of a protein,[13] and to provide additional information for protein identification or confirmation in proteomics. James and coworkers have also successfully applied exopeptidase digestions to obtain sequence information both at the N- and C-termini of tryptic peptides for protein identification.[14,15]

An example of using sequential enzyme digestion to obtain peptide sequence information for protein identification/confirmation is shown in Figure 6.6. Leucine Aminopeptidase M (LAP), which has been successfully used to create N-terminal peptide ladders,[16-18] was used to generate peptide sequence tags from horse cytochrome c tryptic digest. Figure 6.6A is the spectrum of a tryptic digest of horse cytochrome c. Total sample loading was 1.2 femtomoles or 15 picograms of protein. Figure 6.6B shows the spectrum where 2 min of N-terminal digest was performed after tryptic digest. Total sample loading here was 1.4 femtomoles or 17 picograms. Two of the tryptic peptides originating from cytochrome c underwent N-terminal exoproteolytic digestion by LAP. The peptide with average mass 1351.5 Da and sequence TEREDLIAYLK as well as the peptide with average mass 1169.3 Da and sequence TGPNLHGLFGR each lose threonine at their N-terminus, yielding peptides with masses at 1250.4 and 1068.2 Da, respectively. It is not surprising to observe some trypsin autolytic peptides undergo exoproteolytic digestion, since the in-capillary technique employs usually equal or excess amounts of trypsin (compared to protein) to allow for rapid digestion. In Figure 6.6 B, the autolytic peptide with average mass 1154.3 Da (SSGTSYPDVLK) loses two serines at the N-terminus, yielding a peptide with mass 979.5 Da. The results shown in Figure 6.6 demonstrate that additional sequence information is obtainable at the low femtomole or picogram level. The nanoliter technique is therefore an interesting alternative to MS/MS fragmentation techniques in cases where not enough analyte material is available.

Sequential enzymatic digestion was also employed to identify bacterial proteins. Figure 6.7 shows an example. The major protein in fraction #52 was tentatively identified as DNA binding protein HU alpha (P02342, 9535 Da). To confirm the identity, a sequential enzymatic digestion (i.e., trypsin digestion followed by LAP digestion) was performed on this fraction. Panels B and C in Figure 6.7 show sections of mass spectra from in-capillary digestion of this fraction. In the displayed mass range, two peptides underwent exoproteolytic digestion by LAP. If the peptide with monoisotopic mass 958.5 Da is from DNA binding protein HU alpha, it should have a sequence of TGRNPQTGK. A new peptide peak at 857.5 Da due to the loss of the N-

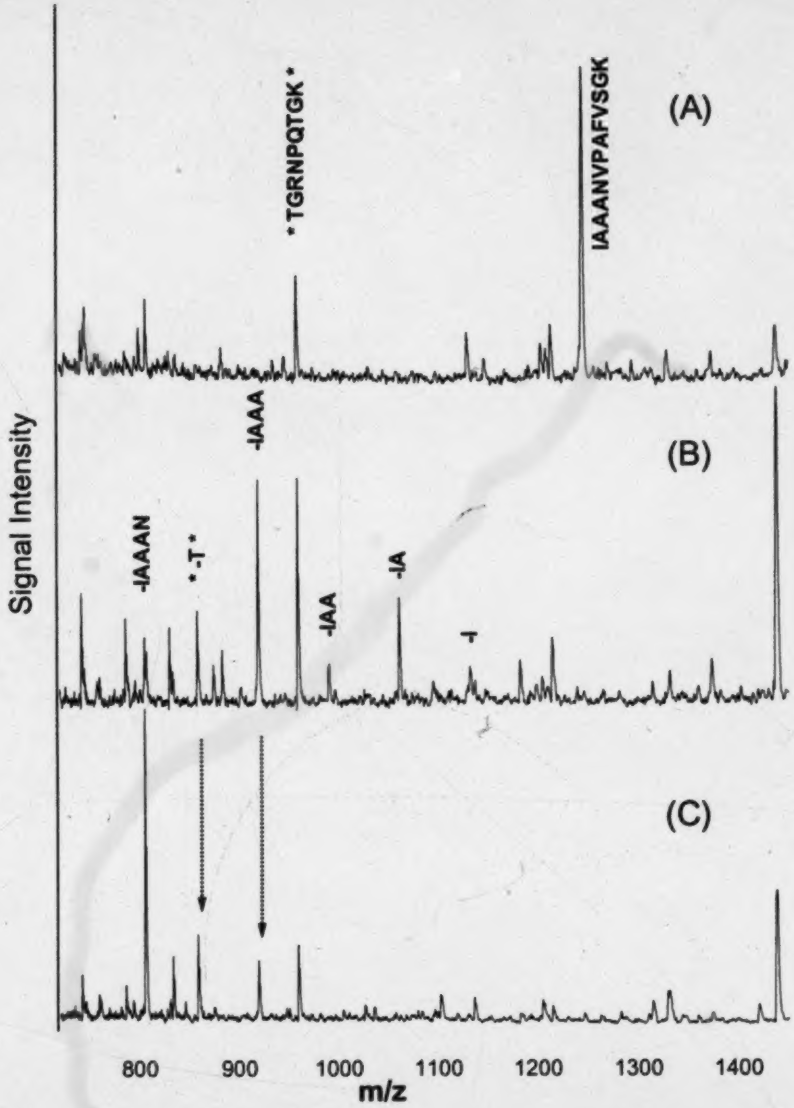

Figure 6.7 Sections of MALDI mass spectra from in-capillary digests of fraction #52 containing DNA binding protein HU alpha. (A) Only trypsin digest (B) Trypsin digest followed by LAP for 5 min. (C) Trypsin digest followed by LAP for 15 min. For each experiment a total volume of ~5 nL was concentrated in ~500 pL portions inside the capillary as described in the Experimental section.

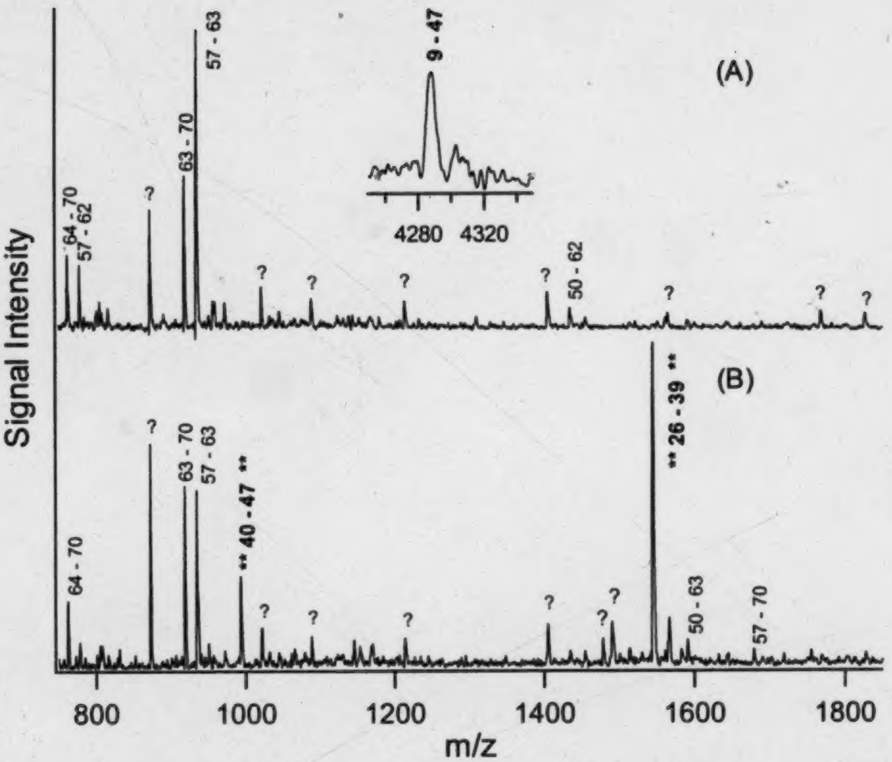

Figure 6.8 MALDI mass spectra of in-capillary tryptic digests of fraction containing 50S ribosomal protein L31. (A) without reduction and alkylation, (B) with reduction and alkylation.

terminal threonine was detected and it did not undergo further digestion by LAP. The peptide with mass 1244.7 Da was tentatively assigned to a sequence of IAAANVPAFVSGK, several N-terminal amino acid losses after LAP treatment were observed. The assigned sequence tags are shown in Figure 6.7C. The exoproteolytic digestion stopped at the V-P bond, since LAP is not capable of cleaving X-P bonds.[22] Note that the signal intensity from the peptide VPAFVSGK (MH^+=804.47 Da) increased with the process of N-terminal digestion, whereas the intensities of all the intermediate N-terminal ladder peptides decreased as shown in Figures 6.7B and 6.7C. The short sequence tag obtained by sequential enzymatic digestion confirmed the identity of the major protein in fraction #52.

The ability to perform multiple reactions in the nanoliter chemstation is very valuable for identifying proteins containing multi-cysteines, since enzymatic digestion usually cannot be effectively performed when proteins are not denatured, consequently, no positive identification can be achieved due to the poor sequence coverage. Protein reduction and alkylation can be easily done inside the capillary.

Figure 6.8 shows the mass spectra of trypsin digests of *E. coli* fraction #26 containing a protein with mass 7867 Da (mass spectrum not shown). Peptide mapping using the data from Figure 6.8A with the molecular weight data identified a top candidate of 50S ribosomal protein L31. However the sequence coverage was only 29%. Moreover, a number of peptide peaks did not match this protein and neither did they match any tryptic autolysis peaks. To increase the confidence in the protein assignment, we examined the digestion data carefully with the assistance of protein structure information contained in the proteome database. 50S ribosomal protein L31 consists of 70 amino acid residues:

23 25 39

MKKDIHP**K**YEEITAS**C**S**C**GNVM**K**I**R**STVGHDLNLDV**C**S**KC**HPFFTG**K**

QRDVATGGRVDRFNKRFNIPGSK

Note that there are four cysteine residues, which may form disulfide bonds, and block the trypsin access to the possible cleavage sites at positions 23 (K), 25 (R), and 39 (K). This assumption was confirmed by the observation of a peptide with m/z at 4287 (See insert of Figure 6.8A), which seems to come from the tryptic peptide from residue 9-47. After reduction and alkylation, two more peptides are detectable covering the sequence from residue 26-39 and 40-47, while the peak at m/z 4287 disappeared (Figure 6.8B). These results illustrate the potential for enhancing the confidence of protein identification by conducting multiple reactions or experiments. With the nanoliter sample handling technique, it is feasible to carry out tens of experiments with only 1 μL sample.

Using the above approach, we are able to identify proteins from two other fractions. The major component in fraction #39 was identified to be 30S ribosomal protein S19 (P02375, 10300Da), and fraction #54 as a mixture of integration host factor beta-subunit (IHF-BETA)(P08756, 10651Da) and DNA binding protein HU alpha (P02342, 9535 Da). A few peaks did not match the identified major proteins in these two fractions. This is expected since in both fractions, some minor protein components were detected by MALDI. However, peptide mass mapping alone cannot positively identify these minor components. It has been found that the majority of the one dimensional HPLC fractions contain multiple proteins with protein numbers generally above five. MALDI protein mass analysis of HPLC fractions reveals over 400 protein components in *E. coli* extract. Enzyme digestion of each fraction can be readily done with the Nanochem Station. However, no positive identification can be achieved by the MALDI spectra of the digests by peptide mass mapping. The sequential digestion protocol is also only applicable to simple protein mixture and many of the spectra from sequential digestion are found to be too complicated to draw useful sequence information for protein identification. These observations are not surprising and indeed they are expected from the analogy performance in identifying proteome displayed in one dimensional gel electrophoresis by using peptide mass mapping. Identification of proteins displayed in 2D-gel is much more successful compared to 1D experiments. Likewise, identification of proteins from HPLC fractions by peptide mass mapping requires the fractions to contain a few proteins, which puts a premier on HPLC separation. Multi-dimensional HPLC is clearly required. Besides, the improvement of separation to reduce the mass spectral complexity for identification, the use of MS/MS to obtain sequencing information on individual peptides should greatly facilitate protein identification. ESI MS/MS has been used widely to identify multiple proteins contained in a 1D-gel spot in proteomics. MALDI MS/MS using quadrupole/time-of-flight mass spectrometer or time-of-flight/time-of-flight instrument has been demonstrated. We envision that MALDI MS/MS, in combination with Nanochem station for protein concentration and digestion, will be an important tool for identifying proteome separated by HPLC.

6.4 Conclusions

We have demonstrated the feasibility of combining analytical HPLC column fractionation with nanoliter chemistry and microspot MALDI TOF analysis. This combination allows the performance of a number of experiments with sample volumes of only a few microliters or less. Such experiments include molecular weight determination, optimization of digestion conditions and multiple chemical or enzymatic reactions. The sequential enzymatic digestion of nanoliter sample volumes yields additional sequence information that allows for more confident protein identification. This technique is therefore an interesting alternative or a complementary technique to MS/MS fragmentation where more sample is usually required or where fragment spectra become complex. Several proteins from *E. coli* extracts are identified after HPLC fractionation. However, using one-dimensional HPLC separation, many individual fractions are found to contain a mixture of several proteins. Although enzyme digestions can be readily performed using the Nanochem station on individual fractions to generate a set of peptides, using the peptide mass alone is found to be difficult to positively identify the proteins from a mixture. This current difficulty with protein identification is expected to be resolved with the use of MALDI MS/MS that is currently becoming commercially available. We envision that the combination of high resolution HPLC fractionation, nanoliter protein concentration and enzyme digestion, and MALDI MS and MS/MS will be a powerful tool for proteome analysis.

6.5 Literatures Cited

1. Wall, D. B.; Kachman, M. T.; Gong, S.; Hinderer, R.; Parus, S.; Misek, D. E.; Hanash, S. M.; Lubman, D. M. *Anal. Chem.* **2000**, 72, 1099.
2. Whittal, R. M.; Keller, B. O.; Li, L. *Anal. Chem.* **1998,** *70*, 5344.
3. Keller, B. O.; Li, L. *J. Am. Soc. Mass Spectrom.* **2001**, 12, 1055.
4. Whittal, R. M.; Li, L. *Anal. Chem.* **1995,** 67, 1950.
5. Keller, B. O.; Li, L. *J. Am. Soc. Mass Spectrom.* **2000,** 11, 88.
6. Gobom, J.; Nordhoff, E.; Mirgoroskaya, E.; Ekman, R.; Roepstorff, P. *J. Mass Spectrom.* **1999,** 34, 105.
7. Technical information on ZipTip Pipette Tips, Millipore Corp., Waltham, MA, USA.
8. Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999,** 20, 3551.

9. Wilkins, M. R.; Gasteiger, E.; Tonella, L.; Ou, K.; Tyler, M.; Sanchez, J. C.; Gooley, A. A.; Walsh, B. J.; Bairoch, A.; Appel, R. D.; Williams, K. L.; Hochstrasser, D. F. *J. Mol. Biol.* **1998,** *278*, 599.

10. Cooks, R. G. *J. Mass Spectrom.* **1995,** *30*, 1215.

11. Biemann, K., In *Methods in Enzymology: Mass Spectrometry*, McCloskey, J. A., Ed., Vol. 193, Academic Press: San Diego, CA, USA, **1990**, pp. 455.

12. Spengler, B. *J. Mass Spectrom.* **1997,** *32*, 1019-1036, and references therein.

13. Doucette, A.; Li, L. *Joint-issues of Proteomics and European Journal of Mass Spectrometry*, **2001**, in press.

14. Staudenmann, W.; Hatt, P. D.; Hoving, S.; Lehmann, A.; Kertesz, M.; James, P. *Electrophoresis* **1998,** 19, 901.

15. Korostensky, C.; Staudenmann, W.; Dainese, P.; Gonnet, G.; James, P. *Electrophoresis* **1998,** 19, 1933.

16. Thiede, B.; Liebold-Wittman, B.; Bienert, M.; Krause, E. *FEBS Letters* **1995**, 357, 65.

17. Woods, A. S.; Huang, A. Y. C.; Cotter, R. J.; Pasternack, G. R.; Pardoll, D. M.; Jaffee, E. M. *Anal. Biochem.* **1995**, 226, 15.

18. Schriemer, D. C.; Yalcin, T.; Li, L. *Anal. Chem.* **1998,** 70, 1569.

Chapter 7

Identification of *E. coli* Proteins and Protein Fragments by MALDI MS and Capillary LC MS/MS

7.1 Introduction

In the previous two chapters, we demonstrated the methodology for identifying proteins from *E. coli* cell extract by one-dimensional HPLC separation, followed by enzyme digestion and MALDI peptide mass mapping. Limited success was achieved due to the fact that most fractions collected by one-dimensional HPLC separation contain more than two protein components. Optimization of the separation conditions, such as using a two-dimensional separation technique, will certainly result in a greater number of fractions that contain only one or two proteins. However, a robust, automated two-dimensional separation technique has not been fully developed. An alternative approach to increase the possibility of identifying proteins is to generate peptide structure information, in addition to their masses, for protein identification. Peptide structure information can be obtained by using tandem mass spectrometry or MS/MS. The integration of liquid chromatography with MS/MS has been proved to be a powerful technique for identifying proteins from protein mixtures.[1-4] LC MS/MS has been used to identify proteins directly from the digest of crude cell lysates. In most cases, protein identification by MS/MS is based on the use of fragment ion spectra from several peptides. However, positive protein identification can be obtained on the basis of a single peptide fragment ion mass spectrum providing the spectral quality is very high (i.e., many different types of fragment ions are generated and well matched with the theoretical spectrum in the database).

In this chapter, capillary LC ESI MS/MS was used to identify proteins in HPLC fractions from an *E. coli* cell extract. The main objective of this work was to understand the origin of the protein masses in the HPLC fractions that were detected by MALDI MS. As indicated in previous chapters, many protein peaks detected by MALDI have masses that do not match with any molecular masses of proteins in the known proteome database. A large number of proteins were identified. Some have observed molecular masses matching those in the proteome database, whereas most of them have undergone proteolytical process *in vivo*, such as N-terminal cleavage of a methionine residue or a piece of signal peptide. Some of the identified peptides were found to be belonging to large and/or hydrophobic proteins. However, the MALDI MS spectrum did not give the molecular ion mass information corresponding to these large proteins. These proteins are likely the fragments of large precursor proteins predicted from the gene sequences.

7.2 Experimental

The *E. coli* 9637 cells were from the Edgewood RDE Center (Aberdeen Proving Ground, MD, USA). The growth conditions for these cells were the same as those described in the previous chapters. Proteins were extracted by micro probe tip sonication using 0.1% TFA as extraction solvent. About 20 mg lyophilized *E. coli* sample was suspended in 1 mL 0.1% TFA solution in a 1.5 mL siliconized vial. Sonication was done using a Branson Sonifier 450 (VWR Scientific, Bridgeport, NJ). The duty cycle was set to 70% and the output control was set to 3. The *E. coli* cell suspension was sonicated for 30 s with the vial sitting on dry ice. The crude protein extract was fractionated by reverse phase HPLC using C_8 column as described in Chapter 3. Fractions were collected every minute during the run, and all the fractions were concentrated to about 10 µL shortly after the collection.

Trypsin digestion was performed on each fraction. About 1 µg trypsin was added into each fraction, which was adjusted to pH 8.5 by 1 mM NH_4HCO_3. Digestions were carried out at 37 °C for about 30 min.

The Applied Biosystems Voyager MALDI mass spectrometer (PerSeptive Biosystems, Inc., Framingham, MA) was used in this work to determine the protein molecular masses. A two-layer method was used to prepare the MALDI sample as described in previous chapters.

LC MS/MS analysis of protein digests were performed on a LCQ^{deca} quadrupole ion trap mass spectrometer equipped with a dynamic nanospray source (ThermoFinnigan, San Jose, CA). The dynamic nanospray source was coupled to a Surveyor HPLC system (ThermoFinnigan, San Jose, CA). The pump flow rate of 100 µL/min was reduced to 2 µL/min using an Acurate microflow processor (LC Packings, San Francisco, CA). All separations used a packed capillary column of 15 cm long and 200 µm i.d packed with 5 µm 218MS (C_{18}) beads (Vydac, Hesperia, CA). The HPLC gradient was 0-35% B in 30 min, followed by 35-70% B in 10 min (Solvent A, 0.5% acetic acid in water; B, 0.5% Acetic acid in acetonitrile, v/v). The nanospray tip used was a 50 µm i.d. tip from New Objective (Woburn, MA). During the HPLC separation, the ion trap repetitively surveyed full scan MS over the m/z range of 400-1800 and executed data-dependent MS/MS scans. MS/MS spectra were acquired using a relative collision energy of 30% (LCQ instrumental settings). An isolation width of 2 m/z units was used and recurring ions were dynamically excluded after two MS/MS spectra were obtained. Interpretation of the resulting MS/MS spectra was done by the Sequest software. The *E. coli* proteome database created from a non-redundant protein database, which was downloaded from NCBI (http://www.ncbi.nlm.nih.gov), was used for database searching.

7.3 Results and Discussion

Protein fragments have been identified by Edman sequencing of gel separated *E. coli* proteins.[5,6] The N-terminal sequence tags of these fragments matched the predicted internal region of the genes in the *E. coli* genome. The fragmentations were attributed to *in vitro* artifacts of sample preparation, *in vivo* events, or translation products initiated at the internal sites of the genes. None of these putative cleavage sites matches any known *E. coli* protease recognition sequences, although little is known of the target specificity of *E. coli* proteases.[7-9] In this work, efforts toward the identification of proteins fragments by LC MS/MS will be discussed.

MALDI analysis of the HPLC fractions showed that, in most cases, at least five protein components were detected in a given fraction. However, when the same fraction was digested and then analyzed by capillary LC MS/MS, many peptides were detected and they were found to belong to tens of different proteins. Among these proteins, some were positively identified based on multiple peptide sequences. One might expect that the molecular masses of these proteins should be detected by MALDI MS. But, this was not the case. An example is given in Figure 7.1 and Table 7.1.

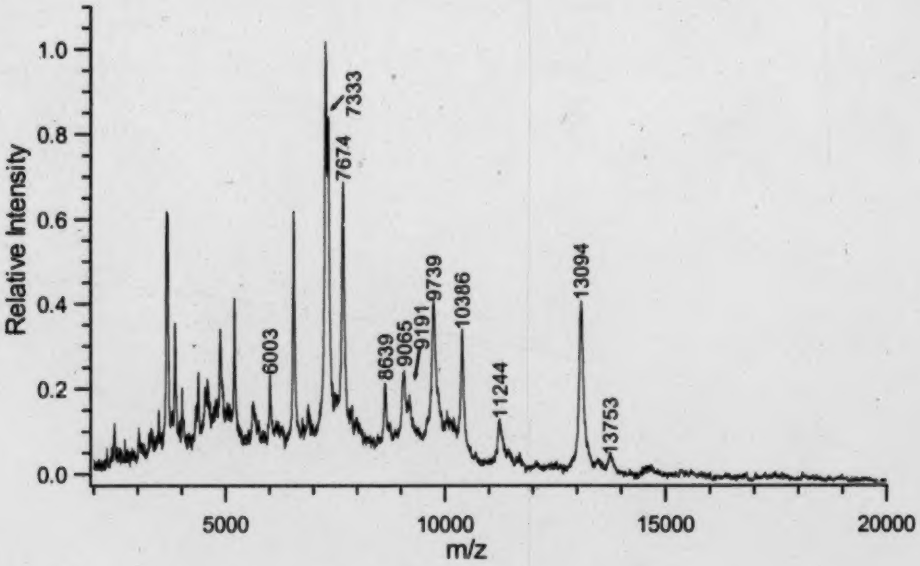

Figure 7.1 MALDI spectrum of fraction #43.

Figure 7.1 shows the MALDI mass spectrum of fraction #43 from HPLC of the *E. coli* extract. The major components detected in this fraction are labelled and the unlabelled peaks with m/z below 7000 are the multiply charged species of these major components. Table 7.1 lists the proteins identified using LC MS/MS analysis of the tryptic peptides generated from this fraction. The proteins are named according to the database list and an accession number corresponding to each protein is also listed. The molecular masses of the proteins from the database are shown. There are 18 proteins identified based on a minimum of two peptide sequences. Another 17 proteins were identified based on a single peptide sequence. Among these 35 proteins, only four proteins have

Table 7.1 Proteins identified from Fraction #43 by LC MS/MS.

Protein	MW (Da) (Observed)	MW (Da) (Predicted)	# of identified peptides	N-terminal processing
UP04 (P39169)		16063	8	
OPPA (P23843)		60899	8	
FKBA (P45523)		28882	8	
CSPA (P15277)	7271	7403	6	-Met
CSPE (P36997)	7332	7463	6	-Met
YBEJ (P37902)		33420	5	
EFTU (P02990)		43314	4	
CSPC (P36996)	7271	7402	4	-Met
MULI (P02937)		8323	4	
HDEA (S30269)	9738	12514	4	-signal peptide
ASG2 (P00805)		36851	4	
HDEB (26605)	9064	12043	3	-signal peptide
ARGT (P09551)		27992	3	
GlNH (P10344)		27190	3	
CH10 (P05380)	10385	10387	3	
CSGA (P28307)	13093	15049	2	-signal peptide
ACP (P02901)	8638	8639	2	
RL16 (P02372)	9190	9190	2	
TIG (P22257)		48193	1	
YCGT (P76015)		39495	1	
YCAC (P21367)		23100	1	
HLPA (P11457)		17688	1	
ALF (P11604)		39147	1	
EFP (P33398)		20580	1	
SPPA (P08395)		67233	1	
TPX (P37901)		17835	1	
YEDF (P31065)	8638	8639	1	
YIBL (P36564)		13696	1	
UP03 (P37903)		16016	1	
ASPA (P04422)		52356	1	
FLGM (P43532)		10340	1	
CLPB (P03815)		95585	1	
YCGT (P76015)		39494	1	
RS20 (P02378)		9684	1	
SYP (P16659)		63733	1	

the database molecular masses matched the peak masses detected in the MALDI spectrum (i.e., 10387, 8639 and 9190). Note that no quantitative information can be obtained from the MALDI or LC MS/MS data. Thus the relative abundances of these proteins in the fraction are unknown. For digestion, the protein in relatively high abundance should generate more peptides than the relatively low abundant ones. However, the more abundant peptide does not necessarily give higher ESI signals. The ionization efficiency for the peptides can be quite different from each other.

Table 7.1 lists three cold shock proteins (CSPs). For these three proteins, the molecular mass difference between the observed peak mass in MALDI and the database mass of the entire protein sequence corresponds to that of methionine (Met), suggesting that the proteins detected in MALDI have the N-terminal methionine removed. For protein HDEA, HDEB and CSGA, a short piece of signal peptide was cleaved. The biological implication of these types of cleaving processing is well understood. For cytoplasmic proteins, the amino-terminal processing model[10,11] predicts that the truncation of N-terminal Met residue by methionine aminopeptidase (PepM) depends on the side-chain length of the second amino acid. When the second amino acid is Ala, Cys, Gly, Pro, Ser, Thr, or Val, the initiator Met is excised. Violations of the model were observed by Edman sequencing of the gel separated *E. coli* proteins,[5] and it was believed that protein structures other than the second amino acid residue are involved in the excision specificity. It has been showed[5] that all initial Met residues are removed when the second residue is Ala and Ser and none of the Met is removed when Val is in the second position. The excise of Met is variable when the second residue is Thr, Gly, or

Pro. The results shown in Table 7.1 seem to support this notion. All three cold shock proteins have lost their initial Met, since they all have Ser in the second position. Ribosome protein S16 has a molecular mass matching the predicted one without cleaving the initial Met. In this case, the second amino acid residue is Val.

Proteins located in the periplasm and outer membrane region are expected to have a signal sequence that helps direct their transport across the inner membrane.[11,12] The signal peptide generally has a positively charged amino-terminal region, a central hydrophobic region, and a carboxy-terminal region. *E. coli* signal peptides are 15 to 30 amino acid long, and they are removed from the protein precursors (i.e., the gene expression products) by the signal peptidase, Lep, after transport through the membrane.[11,13] Therefore, the mature protein has a molecular mass 1500 to 3000 Da lower than that predicted from the genome.

It is interesting to note that some proteins identified by LC MS/MS based on several peptide sequences were not detected in the MALDI spectrum. One possible explanation is the ion suppression effect. Quite a few of these proteins have relatively high molecular masses. thus their ionization could be suppressed by the more easily ionizable low mass proteins. Gel electrophoresis analysis of the cell extracts has revealed the existence of high mass proteins in the extracts. In the HPLC fractionation experiment, we did not treat the sample to remove the high mass proteins. For the C_8 column, high mass proteins from standard proteins such as bovine serum albumin (BSA) and lactoferrin can be separated, albeit at relatively low resolution. Thus the high mass proteins in the cell extract that were injected into the column must have co-eluted with a large number of low mass proteins during HPLC. Since the chromatographic peak from a large protein is expected to be very broad, compared to that from a small protein, the large protein will be eluted out over a long chromatographic time scale. As a result, in a given fraction, the concentration of an individual small protein is likely to be much higher than that from a large protein. Indeed, in several adjacent fractions, the same large proteins are detected by LC MS/MS. In contrast, low mass proteins are generally observed in fractions without much carrier over from one fraction to another. Not surprisingly, MALDI mass spectra did not reveal the molecular ion peaks for large proteins. A good example is for the analysis of protein FKBA (P45523) and YBEJ (P37902) with a molecular mass of 26 and 33 kDa, respectively. Table 7.1 shows that these two proteins are present in this fraction. However, the MALDI spectrum shown in Figure 7.1 does not reveal their molecular ion peaks. However, the presence of the two proteins in the cell extract was confirmed by other experiments. Gel electrophoresis of the cell extract showed two bands in the molecular mass around 26 and 33 kDa.[14] These two bands were positively identified by in-gel digestion, followed by peptide mass mapping or LC ESI MS/MS analysis. The protein molecular masses were also detected by MALDI after the proteins were extracted from the gel bands.[14] Note that direct gel electrophoresis of the HPLC fraction was not possible, because of the low quantity of proteins present in the fraction.

Besides the ion suppression effect, another possible course for not detecting the high mass proteins in MALDI is that only the fragments of the precursor proteins or gene expressed proteins are presented in the cell extract, instead of the intact proteins expected from the gene sequences. Protein fragmentation could have occurred *in vivo* during cell cycle or during protein extraction and isolation. In addition, the proteins observed could be expressed from some internal sites within the genes. This possibility is supported by the fact that, in most cases, only a certain part of the intact protein sequence is covered by the identified tryptic peptides (see below). In addition, there are a number of low mass ions in the MALDI spectra which match with the fragments of larger proteins. Using a software called Paws that was downloaded from http://prowl.rockefeller.edu/, we have examined if there is any possibility to correlate the identified proteins in Table 7.1 and the unidentified low mass species in the MALDI spectrum. By entering a mass of peptide, Paws will search over the entire protein sequence to see if any stretch of sequence whose mass matches with the input mass. In our case, if the observed mass in MALDI matches a part of the sequence which is also covered by the identified peptide sequences, the low mass species we observed in MALDI may be considered as the fragment of the specific protein. Additional supporting evidence is certainly needed to have a conclusive identification.

The identification of protein UP04 (P39196) is shown here as an example of this approach for identifying possible protein fragments present in cell extracts. The amino acid sequence of UP04 is

GLFNFVKDAGEKLWDAVTGQHDKDDQAKKVQEHLNK**TGIPDADKVNIQIADGKATVTGDGLSQEAKEKILVAVGNISGIASVDDQVKTATPATASQFYTVKSGDTLSAISKQVYGNANLYNKIFEANKPMLK**SPDKIYPGQVLRIPEE ,

Its molecular mass is predicted to be 16063 Da. The protein was identified based on the sequences of 7 tryptic peptides, which covered the bolded protein sequence. Table 7.2 lists the identified peptides with their statistical scores. Using Paws, it is found that the underlined protein sequence has a molecular mass of 13755 Da, which closely matched the small peak shown in Figure 7.1 with m/z at 13753. Additional information is required to confirm that the observed MALDI peak is actually a fragment of this UP04 protein. One method to be explores in the future, is the use a preparative column to collect a sufficient amount of proteins in individual fractions. We will then take a portion of the fraction for gel electrophoresis, followed by in-gel digestion of the observed bands for MS protein identification. For example, if a band corresponding to the mass of ~13755 Da should be observed in the gel image of the fraction and identified as UP04, this would provide a confirmation of the protein being the fragment of UP04.

Table 7.2 Tryptic peptides matched protein UP04-ECOLI.

Protein	Identified peptides	# b/y	Xcorr	dCn
UP04_ECOLI (P39169)	ILVAVGNISGIASVDDQVK (1899 Da)	13/13	5.75	0.68
	TGIPDADKVNIQIADGK (1755.5 Da)	9/12	4.65	0.56
	SGDTLSAISK (979.3 Da)	8/9	3.42	0.41
	ATVTGDGLSQEAK (1277.3 Da)	8/9	3.29	0.46
	TATPATASQFYTVK (1485.8 Da)	7/10	2.97	0.60
	QVYGNANLYNK (1284.3 Da)	6/8	2.93	0.55
	IFEANKPMLK (1191.2 Da)	8/8	2.82	0.38

Another example is for protein EFTU (P02990) that was identified from the sequence information of four peptides, as shown in Table 7.1. The MS/MS spectra of these peptides are shown in Figure 7.2. The predicted protein has a molecular mass of 43 kDa with 393 amino acid residues. However, the identified peptides only cover the part of sequence from amino residue 9-75. The molecular mass from residue 1-104 is 11216 Da, which might be related to the peak with m/z at 11244. The difference of the masses

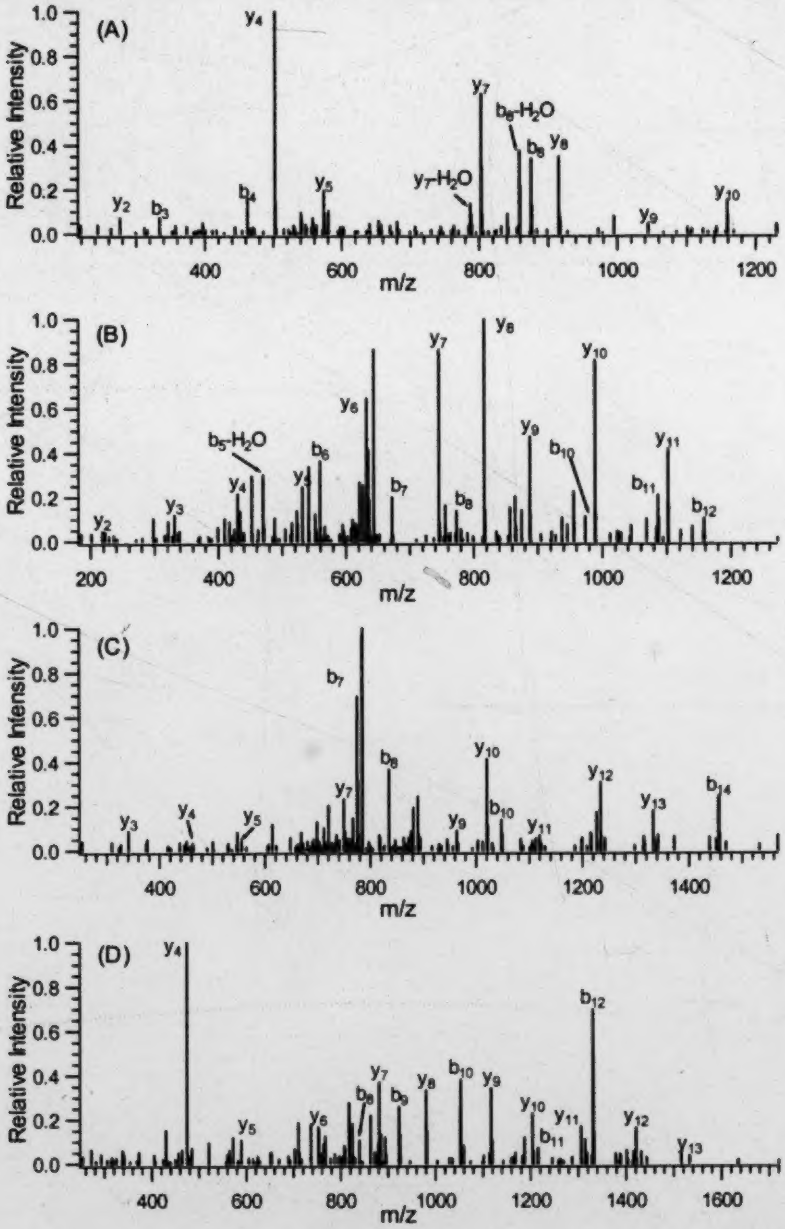

Figure 7.2 MS/MS spectra of tryptic peptides from EFTU_ECOLI. (A) AFDQIDNKPEEK (position 46-57), (B) TTLTAAITTVLAK (position 26-38), (C) GITINTSHVEYDTPTR (position 60-75), (D) TKPHVNVGTIGHVDHGK (position 9-25).

could be due to the oxidation of the methionine residues (position 91 and 98) during sample preparation. Note that the accuracy for mass measurement is about 0.05%.

The above discussion only suggests that some of the unidentified molecular ions in the MALDI spectra could be the fragments of larger proteins. In some cases, posttranslational modifications may also be involved. More experiments, such as gel electrophoresis of individual fractions, are required to confirm the identification of these species.

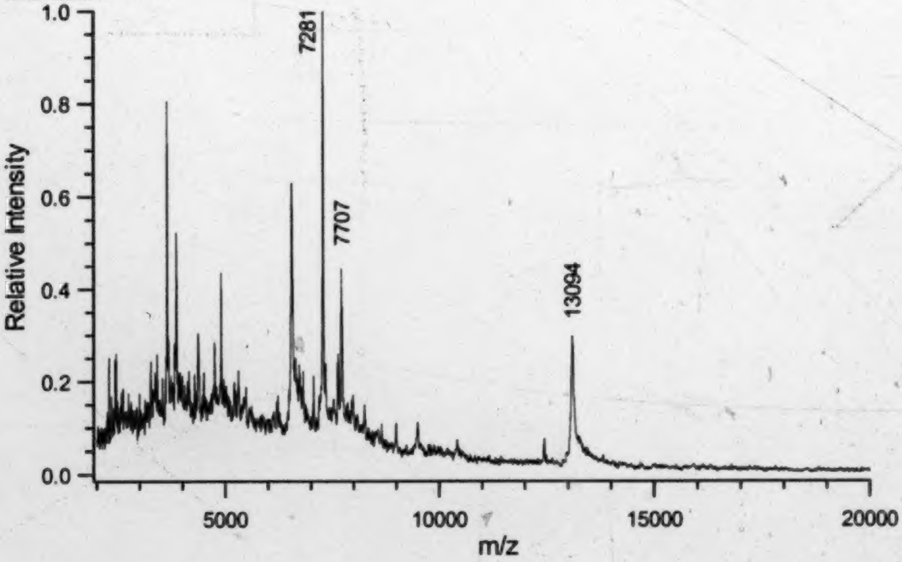

Figure 7.3 MALDI spectrum of fraction #32.

In one fraction, our effort did positively identify a protein fragment. Figure 7.3 shows the MALDI spectrum of fraction #32. There are several proteins present in this fraction, with three major ones at m/z's 13094, 7707 and 7281.

Using LC MS/MS, DNA-binding protein H-NS (Histone-like Protein HLP-II)(Protein H1)(Protein B1) (HNS_ECOLI) (P08936) was identified based on the fragmentation patterns of 6 tryptic peptides. Figure 7.4 shows two representative MS/MS spectra. The protein has the following amino acid sequence: MSEALKILNNIRTLRAQARECTLETLEEMLEKLEVVVNERREEESAAAAEVEERTRKLQQYREMLIADGID **PNELLNSLAAVK** ***SGTKAKRAQRPAKYSYVDENGETKTWTGQGRTPAVIKKAMDEQGKSLDDFLIKQ***

The italic bolded sequence was covered by the MS/MS results. The molecular mass of this protein is 15540 Da, which was not found in the MALDI spectrum (Figure 7.3). Using the Paws program, it is found that the underlined protein sequence has the molecular mass of 7280 Da, matching one major component in Figure 7.3. The molecular ion with m/z at 7281 is most likely the fragment from HNS_ECOLI. The identification became confident when an unexpected tryptic peptide PNELLNSLAAVK (Figure 7.5) was identified. In this case, Sequest database searching was done without any enzyme type constraint. Nine y ions and eight b ions matched the *in silico* fragmentation pattern of the peptide PNELLNSLAAVK, showing a strong correlation between the MS/MS spectrum and the identified peptide sequence. Note the fragmentation site is between Asp (D) and Pro (P) residue. It is found that in dilute acid conditions, aspartyl peptide bonds tend to be more rapidly hydrolyzed than other

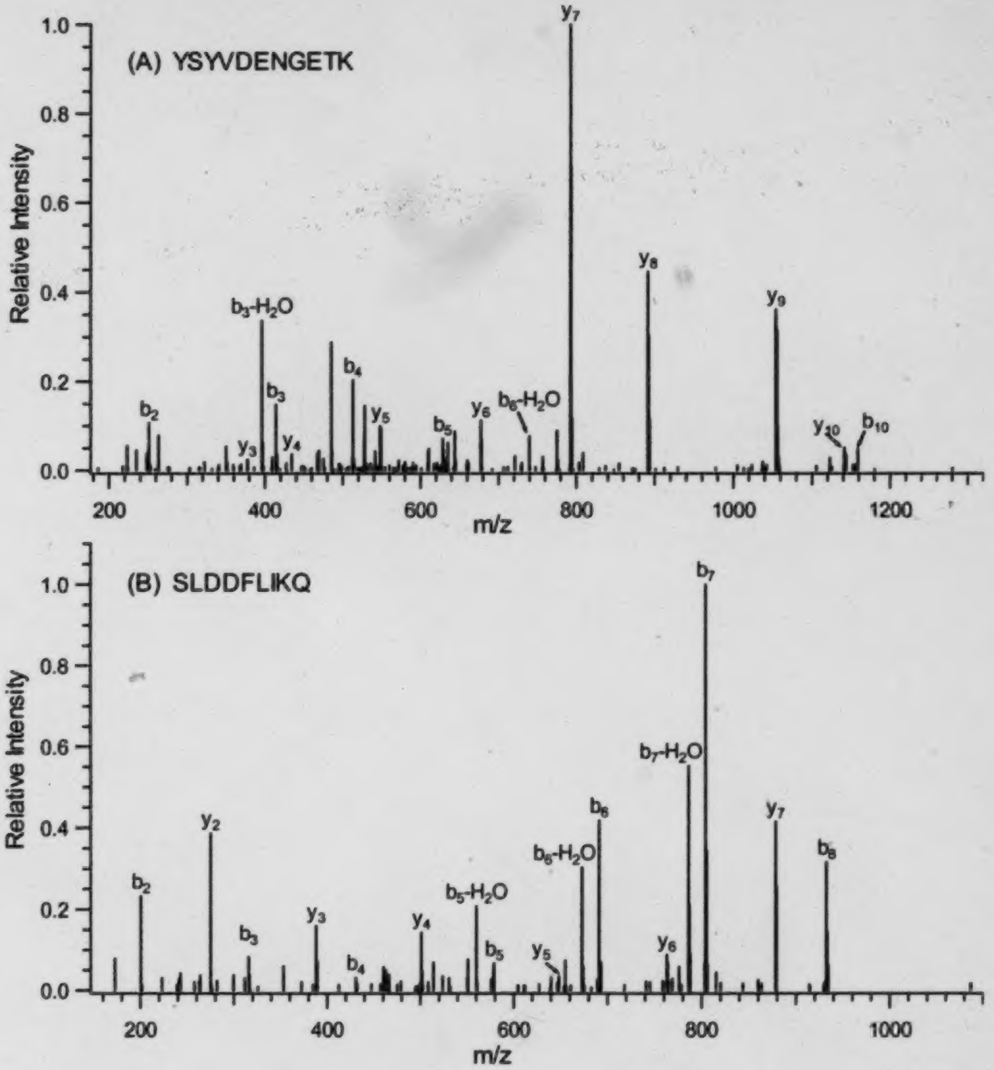

Figure 7.4 MS/MS spectra of two tryptic peptides from HNS_ECOLI.

aminoacyl peptide bonds, particularly for aspartyl proline bonds. This is due to the neighboring group effects caused by the proximity of the side-chain carboxyl group on amino acid residue Asp to the α-carboxyl peptide bond.[16] On the basis of this notion, the fragmentation of HNS_ECOLI most likely occurred during protein extraction and isolation, since dilute TFA aqueous solution was used in both steps.

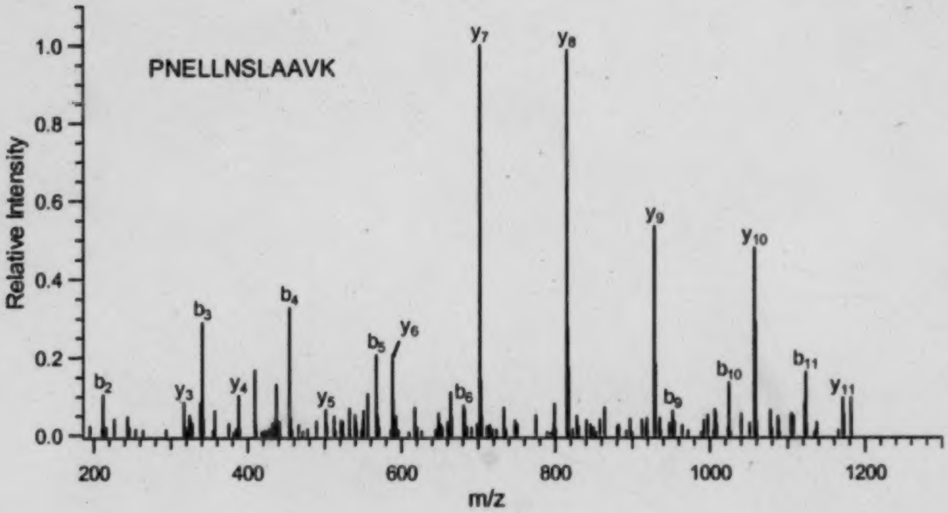

Figure 7.5 MS/MS spectrum of an unexpected tryptic peptide from HNS_ECOLI.

The other two major components in this fraction were identified as YAHO_ECOLI (P75694) with a predicted molecular mass of 9895 Da and CSGA_ECOLI (P28307) with a predicted molecular mass of 15049 Da. Both proteins have lost a short piece of N-terminal signal peptide, resulting in the molecular ion signals with m/z at 7707 and 13094, respectively. The identification is based on several peptide sequences for both proteins as shown in Table 7.3. The statistical scores shown indicate the high confidence of identification. Note that protein CSGA was also detected in the MALDI spectrum of fraction #43, and it was also identified from that fraction based on several peptide MS/MS patterns. This finding indicates that this protein exists in both fractions. During direct MALDI analysis of the crude *E. coli* cell extracts, it is often found that the molecular ion with m/z at 13094 and its multiply charged species can dominant the spectra. This protein is detected in many HPLC fractions, indicating that this protein is likely in very high abundance in the crude extracts.

Table 7.3 Tryptic peptides matched YAHO_ECOLI and CSGA_ECOLI from Fraction #32.

Protein identity	**Peptides identified**	**#b/y**	**Xcorr***	**DelCn***
YAHO_ECOLI	GADVLVLTSGQTDNKIHGTANIYK	11/8	6.04	0.61
(P75694)	IGDISTSNEMSTADAKEDLIK	7/10	5.92	0.59
(9895/7706 Da)	GADVLVLTSGQTDNK	11/10	5.30	0.70
	IGDISTSNEMSTADAKEDLIKK	5/15	4.31	0.67
	IGDISTSNEMSTADAK	12/11	3.76	0.62
	AEFEKVESQYEK	9/9	3.46	0.57
	IHGTANIYK	6/7	2.61	0.51
CSGA_ECOLI	NSDLTITQHGGGNGADVGQGSDDSSIDLTQR	16/20	8.46	0.69
(P28307)	QFGGGNGAAVDQTASNSSVNVTQVGFGNNATAHQY	18/18	7.14	0.69
(15049/13093 Da)	GFGNSATLDQWNGK	8/7	3.10	0.58

Table 7.4 lists the proteins identified from all the HPLC fractions. Only those proteins identified based on two or more peptide fragmentation patterns are included in this table. It is clear from Table 7.4 that the situation for fraction #43 represents the general trend. Most of the identified proteins are involved in well known proteolytic processing such as N-terminal truncation of Met or a signal peptide. Many proteins, especially large proteins, were not detected by the MALDI analysis of the individual fractions. Besides the large proteins, some hydrophobic proteins, such as major outer membrane proteins, are also identified. This is surprising since these membrane proteins should not be extracted by 0.1% TFA aqueous solution. One possible explanation is that only the hydrophilic parts of the proteins were knocked out of the protein during sample

preparation or during the cell cycle. For example, outer membrane protein A (OMPA) has 346 amino acid residues and its sequence is as follows:

MKKTAIAIAVALAGFATVAQAAPKDNT***WYTGAKLGWSQY***HDTGFINNNGPTHE***NQLGAGAFGGYQV***NPY***VGFEMGYDWLGRM***PYKGSVENGAYK***AQGVQLTAKGY***PITDD***LDIYTRLGGMVWRA***DTKSNVYGKNHDTG***VSPVFAGGVEYAI***TPE***IATRLEYQWTNN***IGDAHTIGTRPD***NGMLSLGVSYRF***GQGEAAPVVAPAPAPAPEVQTKHFTLKSDVLFNFNKATLKPEGQAALDQLYSQLSNLDPKDGSVVVLGYTDRIGSDQGLSERRAQSVVDYLISKGIPADKISARGMGESNPVTGNTCDNVKQRAALIDCLAPDRRVEIEVKGIKDVVTQPQA

The italic bolded sequences represent the transmembrane domains. The sequences identified by the tryptic peptides cover the underlined sequence, suggesting the identified protein might be a fragment of this hydrophobic protein.

7.4 Conclusions

A large number of *E. coli* proteins were identified by LC MS/MS analysis of the tryptic digests of the HPLC fractions. By comparing the observed molecular masses with the predicted ones, it is found that most of these proteins have involved in post-translational modification, especially proteolytic processing. Some of the high mass proteins were identified and the absence of their molecular mass information in the MALDI spectra is attributed to the ion suppression effect or protein fragmentation. Protein fragmentation for some large proteins was suggested based on the molecular mass information and the peptide sequence coverage observed. Among them, one was positively identified and this fragment protein was found to be from an *in vitro* process. The identification of some hydrophobic membrane proteins further suggests the possibility of protein fragmentation during sample preparation or cell cycle. The biological significance of this observation will be investigated in the future.

It is worthy noting that the results shown in this work demonstrate the limitation of using the current public proteome database for bacteria identification. The protein fragment information is not included in the public proteome database. However, the masses of these fragments can still be specific for the bacterium, and thus be useful for the purpose of bacteria identification. A mass database generated by the mass spectrometric method (see Chapter 3) includes these fragments and should be more useful for bacteria identification.

Table 7.4 Proteins identified from HPLC fractions by LC MS/MS.

Protein	M. W. (Da, Observed)	M.W. (Da, Predicted)	# of peptides identified	N-terminal processing
YMDF (P56614)	5752	5883	4	-Met
YCIG (P21361)	5871	6002	2	-Met
RS22 (P28690)	5096	5096	5	
RL33 (P02436)	6255	6372	5	-M and methylated
RL32 (P02435)	6315	6446	3	-Met
YDCH (P46135)	6338	6470	3	-Met
YDFY (P77695)	6679	6679	3	
CSRA (P31803)		6856	2	
RL29 (P02429)	7274	7273	3	
YAIA (P08366)		7281	3	
CSPC (P36996)	7271	7402	6	-Met
CSPA (P15277)	7271	7403	6	-Met
CSPE (P36997)	7333	7463	6	-Met
RL31(P02432)	7867	7871	5	-4H (disulfide bonds)
MULT (P02937)		8323	5	
YJBJ (P32691)		8325	3	
YEDF (P31065)	8638	8639	2	
ACP (P02901)	8638	8640	3	
CHAB (P39162)	8814	8945	5	-Met
RL28 (P02428)	8876	9006	2	-Met
RL27 (P02427)	8994	9124	3	-Met
GLR3 (P37687)		9137	2	
RS16 (P02372)	9190	9191	7	
DBHB (P02341)	9226	9226	6	
YCIN (P46132)		9386	2	
DINJ (Q47150)		9406	3	
DBHA (P02342)	9535	9535	5	
RS20 (P02378)	9555	9684	5	-Met
HDEA (S30269)	9738	12514	4	-signal peptide
YAHO (P75694)	7706	9895	5	
RPOZ (P08374)	10105	10237	7	
YIHO (P32126)		10273	2	
FLGM (P43532)		10341	2	
CH10 (P05380)	10386	10387	5	
RS19 (P02375)	10300	10430	5	-Met
YDHR (P77225)		11288	2	
RL24 (P02425)	11185	11316	12	-Met
IHFA (P06984)	11221	11354	9	-Met
PSPE (P23857)	9426	11475	4	-signal peptide
TH10 (P00274)		11807	5	
YBAB (P17577)		12015	3	
HDEB (P26605)	9065	12043	3	-signal peptide
YNFD (P76172)	8449	12139	3	-signal peptide
RL7 (P02392)		12295	3	
YAJD (P19678)		12589	3	
RL18 (P02419)	12769	12770	8	
YFIA (P11285)		12785	3	
YBGS (P75758)	10463	12872	3	-signal peptide
YDHD (P37010)		12879	2	
YNFB (P76170)	9977	12909	2	-signal peptide

YJGF (P39330)		13612	4	
YIBL (P36564)		13696	4	
RS11 (P02366)		13845	3	
YGIW (P52083)	11976	14011	7	-signal peptide
C562 (P00192)	11780	14061	9	-signal peptide
YFID (P33633)		14284	6	
YQJC (P42616)		14466	4	
RL11 (P02409)		14875	3	
CSGA (P28307)	13093	15049	3	-signal peptide
CSGF (P52104)		15056	2	
YAEH (P37048)		15096	2	
YHCB (P39436)		15239	2	
HNS (P08936)	7280	15540	7	fragment
SLYB (P55741)		15602	4	
RL9 (P02418)		15769	2	
UP12 (P39177)		15935	5	
RL13 (P02410)		16019	2	
UP04 (P39169)		16063	9	
YHHA (P23850)	14735	16624	8	-signal peptide
UP18 (P45502)		16872	2	
RS5 (P02356)		17603	9	
SODC (P53635)		17681	3	
HLPA (P11457)	15693	17688	8	-signal peptide
SPY (P77754)		18199	9	
PTGA (P08837)		18251	4	
PAL (P07176)		18824	5	
RL6 (P02390)		18904	5	
DCRB (P37620)		19787	5	
CYPH (P20752)		20431	3	
RRF (P16174)		20639	4	
OSMY (P27291)	18160	21074	7	-signal peptide
SODF (P09157)		21266	2	
FKBB (P39311)		22216	4	
RL3 (P02386)		22244	2	
DEDD (P09549)		22938	2	
YCAC (P21367)		23100	2	
RSEA (P38106)		24321	3	
RS2 (P02351)		26744	2	
GLNH (P10344)		27190	6	
ARGT (P09551)		27992	3	
PMG1 (P31217)		28556	2	
FKBA (P45523)		28882	8	
FLIY (P39174)		29039	3	
SUCD (P07459)		29777	4	
RL2 (P02387)		29860	5	
KDSA (P17579)		30833	2	
MALM (P03841)		31943	2	
MDH (P06994)		32337	2	
YBEJ (P37902)		33420	5	
YDGH (P76777)		33903	3	
G3P1 (P06977)		35532	2	
YHDW (P45766)		33426	2	
DGAL (P02927)		35713	11	
ZIPA (P77173)		36433	2	
ASG2 (P00805)		36851	7	
YHDW (P45766)		37020	2	

OMPA (P02934)	37201	10
OMPF (P08366)	39333	3
FTSZ (P06138)	40297	4
OMPC (P06996)	40368	3
SUCC (P07460)	41393	3
ACRA (P31223)	42196	4
EFTU (P02990)	43314	4
ENO (P08324)	45655	3
DFP (P24285)	46301	2
PURA (P12283)	47345	2
TIG (P22257)	48193	3
SYS (P09156)	48414	2
FUMC (P05042)	50489	3
FLIC (P04949)	51295	5
ASPA (P04422)	52356	5
TOLC (P02930)	54014	2
CH60 (P06139)	57269	2
PPCK (P22259)	59643	5
DPPA (P23847)	60294	3
OPPA (P23843)	60899	11
YFJL (P52127)	62006	2
DNAK (P04475)	69115	7
MAO2 (P76558)	82417	2
ACO2 (P36683)	93498	2
ADHE (P17541)	96127	2
IF2 (P02995)	97350	3

7.4 Literatures Cited

1. Griffin, P. R.; Coffman, J. A.; Yates, III, J. R. *Int. J. Mass Spectrom. Ion Proc.* **1991**, 111, 131.

2. Davis, M. T.; Stahl, D. C.; Hefta, S. A.; Lee, T. D. *Anal. Chem.* **1995**, 67, 4549.

3. Link, A. J.; Carmack, E.; Yates, III, J. R. *Int. J. Mass Spectrom. Ion Proc.***1997**, 160, 303.

4. McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin. D. *Anal. Chem.* **1997**, 69, 767.

5. Link, A. J.; Robison, K.; Church, G. M. *Electrophoreis*, **1997**, 18, 1259.

6. Wasinger, C. V.; Humphery-Smith, I. *FEMS Microbio. Lett.* **1998**, 169, 375.

7. Maurizi, M. R. *Experientia*, **1992**, 48, 178.

8. Lazdunski, A. M. *FEMS Microbio. Rev*, **1989**, 63, 265.

9. Miller, C. G. in: Neidhardt, F. C.; Curtiss, R.; Gross, C. M.; Schaechter, M.; Umbarger, H. E. (Eds), *Escherichia coli and Salmonella: Cellular and Molecular Biology*, ASM Press, Washington, DC 1996, pp 938-954.

10. A. Ben-Bassat, A.; Bauer, K.; Chang, S. Y.; Myambo, K.; Boosman, A.; Chang, S. J. *J. Bacteriol*. **1987**, 169, 751.

11. Murphy, C. K. Beckwith, J. in: Neidhardt, F. C.; Curtiss, R.; Gross, C.; Ingraham, J. L.; Lin, E. C.; Low, K. B.; Magasanik, B.; Riley, M.; Schaechter, M.; Umbarger, H. E. (Eds.), *Escherichia coli and Salmonwlla: Cellular and molecular Biology*, ASM Press, Washington, DC, 1996, pp 967-978.

12. Hirel, P. H.; Schmitterm, J. M.; Dessen, P.; Fayat, G.; Blanquet, S. *Proc. Natl. Acad, Sci*. USA, **1989**, 133, 17.

13. Von Heijne, V. *Biochem. Biophys. Acta*, **1988**, 947, 307.

14. Wang, Z.; Zheng, J.; Li, L. *J. Mass Spectrom*. Submitted.

15. Young, C. C.; Bernlohr, R. W.; *J. Bacteriol*, **1991**, 173, 3096.

16. Allen, G. *Protein*, JAI Press, London, England, **1997**, 1, 14.

Chapter 8

Conclusions and Future Work

The goal of this work is to develop a fast, sensitive and reliable approach for bacterial identification using MS techniques. A set of protein masses, directly detected from crude bacterial cell extracts or whole cells by MALDI, are used as biomarkers to search against a bacterial protein mass database generated by MS methods. The possible bacteria candidates can be retrieved from the database with statistic scores. This bacterial identification approach is similar to protein identification based on a set of tryptic peptide masses from a protein (i.e., peptide mass mapping). The entire genome of a bacterium consists of many genes from which many proteins are expressed. Detection of a subset of these expressed proteins should be adequate for unique identification of the bacteria. As in peptide mass mapping for protein identification, the confidence of bacterial identification based on protein mass analysis increases with the quality of the mass spectra data and the quality of the bacteria database. This work mainly focuses on developing methods to improve the MALDI technique for bacterial protein analysis and understanding issues related to the creation of protein mass database tailored for bacterial identification.

In Chapter 2, several experimental factors related to mass spectral reproducibility in direct MALDI analysis of proteins and peptides from crude bacterial cell extracts were systematically investigated. With regard to sample preparation, it has been demonstrated that the solvent composition in preparing the MALDI matrix/sample solution and the salt content in the sample have a significant effect on mass spectral pattern. On the issue of protein extraction, we find that the solvent suspension method provides a rapid means of extracting peptides and proteins from bacterial samples. However, different mass spectra may be obtained using different protein extraction processes. The type of extraction solvent and the pH of the aqueous solution used for extraction can have a major impact on the observed spectra. Using the same optimized sample extraction/preparation strategies, it is found that reproducible mass spectra can be obtained, suggesting that the technique has the potential to be a valuable bacterial identification tool. It is also found that despite the significant variation of mass spectral patterns resulting from minor changes in experimental conditions, many peaks are consistently detected from a specific bacterium. These "conserved" peaks represent the ones that have the highest potential for use as biomarkers for bacterial identification.

In Chapter 3, we proposed that a bacterial protein mass database specifically tailored for bacterial identification can be generated by MS methods. Bacterial identification on the basis of searching a set of protein masses against such a database will not be affected by the variations in protein mass spectral pattern. Different sets of protein masses obtained under different experimental conditions should retrieve the same bacterium by searching the bacterial protein mass database, since the sets of protein masses should always reflect the bacterial genome. To achieve a confident identification, it is critical to have a comprehensive and reliable protein mass database. Proteome database in the public domain provides protein sequence as well as their molecular mass information, but it cannot be directly used for the purpose of bacterial identification. This is due to the fact that only a few bacterial species have relatively complete proteome database. Moreover, the protein masses in the proteome database are mostly derived from their genome translated protein sequences and are not experimentally confirmed. In reality, most proteins have involved in some kinds of *in vivo* processing after translation (i.e., post-translational modification such as proteolytic processing). Proteins can also be modified or processed *in vitro* during sample manipulation. A protein mass database for bacterial identification should take into account of such information. We believe that protein mass database can be easily created by MS methods. Issues related to the database creation are discussed in detail in Chapter 3. Preliminary results have demonstrated that such a database is potentially very useful for confident bacterial identification.

In Chapter 4, we focused on protein extraction efficiency and how this will affect direct MALDI analysis of crude bacterial cell extracts. Confident bacterial identification will not only rely on the completeness and reliability of the protein mass database, but also depend on the number of proteins that can be detected by MS analysis of the extracts. The detection of a large set of proteins in a wider mass window will certainly increase the confidence of bacterial identification. This can be achieved by using a more efficient protein extraction procedure (i.e., probe tip sonication instead of vortexing) and optimized MALDI sample preparation.

In Chapters 5 and 6, we describe the research efforts on the identification of bacterial proteins. Results from Chapters 2-4 indicate that quite a number of protein masses detected by MS from the *E. coli* extract cannot match any proteins in the known proteome database in the public domain. Knowing the origin of these proteins is important in validating the protein mass database created by the MS techniques. In Chapter 5, using HPLC fractionation, followed by multiple-enzyme digestion and MALDI TOF MS analysis, several proteins, including some post-translational modifications and protein degradation products, are successfully identified from an *E.*

coli extract. These results provide direct evidence that some of the protein masses detected by MS techniques, particularly those involved in post-translational modifications, were not counted in the proteome database. The mass spectrometric approach presented herein should be very useful to provide information on protein modifications that can be incorporated into the database, thereby enhancing the utility of the current proteome database for bacterial identification as well as biological applications (e.g., functional studies). To address the sensitivity issue in using protein mass mapping method for protein identification, in Chapter 6, a technique for pre-concentrating protein solutions inside a capillary tube, followed by chemical and enzymatic reactions and microspot MALDI analysis was presented. This method allows multiple experiments to be carried out from a very small volume of diluted protein sample. This makes it possible to identify bacterial proteins fractionated by an analytical HPLC column. The improved sensitivity with this technique in protein identification would also be very valuable for other proteomics projects involving a limited supply of cells.

Despite of the efforts presented in Chapters 5 and 6, only limited success was achieved in protein identification from the fractions collected by one-dimensional HPLC separation. It is found that most of the HPLC fractions contain more than five protein components and peptide mass mapping alone cannot provide adequate information for positive identification. In Chapter 7, LC MS/MS was used to achieve extensive protein identification from the relatively complex protein fractions. A large number of proteins were identified and their predicted molecular masses were compared with those detected by MALDI analysis of the fractions. It is found that most of these proteins have undergone post-translational processing, such as the cleavage of N-terminal methionine and signal peptides, resulting in the deviation of the observed molecular masses from that predicted from the genome sequence. In addition, the protein components detected by MS might be the fragmentation products of large proteins - the fragmentation could have occurred *in vivo* or *in vitro*. This finding implies that a protein mass database specifically tailored for bacterial identification is necessary, since protein fragmentation information, particularly those fragmentation during sample preparation, will not likely to be included into the proteome database. However, the masses of these fragments might still be specific for the bacterium and thus be useful for the purpose of bacterial identification. A mass database generated by the mass spectrometric method will include this information and can be potentially used as an alternative and complementary database for bacterial identification.

Work described in Chapters 2-7 demonstrates the improved methods for MALDI analysis of bacterial proteins as well as the importance of creating protein mass database by MS for bacterial identification. Future work will continuously focus on developing a protocol for database creation based on the preliminary results shown in Chapter 3. To establish a comprehensive and reliable bacterial protein mass database, bacteria grown under various conditions (i.e., different growth times and different growth media) and analyzed under different experimental conditions (i.e., different extraction methods and different MS techniques) must be examined. Those consistently observed protein masses under various conditions will be included in the mass database. The database creation based on this strategy is current underway for several bacteria of interest. To apply the database for identification of bacteria in the particles collected from the atmosphere, mass database containing several 'background' bacteria in ambient air will be constructed.

Identification of bacteria of interest against the 'background' species will be studied and demonstrated.

Another direction in bacterial identification using the database searching approach is to examine the feasibility of generating MS/MS spectra of intact proteins and including them into the mass database. This will provide another level of information (protein sequence) to increase the specificity in bacterial identification. The inclusion of protein sequence information will be very important, since protein masses detected by MS techniques are mostly in the mass range of 2-20 kDa. Even with optimized protein extraction procedures and sample preparations, a high percentage of protein masses in the databases would still be expected to be in the low mass range. Many proteins congested in a relatively narrow mass window will potentially result in uncertainty in bacterial identification. This uncertainty will become prominent when dealing with bacterial mixtures, which will most likely be the case for real world bacterial sample. The fragmentation of multiply charged protein ions can be done in an ion trap mass spectrometer using low energy CID with ESI. The fragmentation patterns of different proteins with the same molecular mass will be very different. One study has demonstrated that nine species of cytochrome c can be differentiated using the CID spectra of the intact protein ions.[1] Thus, it is well worth examining the possibility of including protein fragmentation information into the mass database to increase the specificity for bacterial identification.

Literature Cited

1. Smith, R. D.; Loo, J. A.; Barinaga, C. J.; Edmonds, C. G.; Udseth, H. R. *J. Am. Soc. Mass Spectrom.* **1989**, 1, 53.